Probabilities and Statistics Notes

Zhi-Qiang Zhou

November 27, 2021

Contents

Preface

Most contents are from Bain & Engelhardt (2000).

Chapter 1

Probability

1.1 Notation and Terminology

Definition 1.1.1: Sample Space

The set of all possible outcomes of an experiment is called the **sample space**, denoted by S.

Note that one and only one of the possible outcomes will occur on any given trial of the experiments.

Definition 1.1.2: Discrete Sample Space

If a sample space S is either finite or countably infinite, then it is called a **discrete sample space**.

Definition 1.1.3: Event

An **event** is a subset of the sample space S. If A is an event, then A has occurred if it contains the outcome that occurred.

Definition 1.1.4: Elementary Event

An event is called **elementary event** if it contains exactly one outcome of the experiment.

Definition 1.1.5: Mutually Exclusive

Two events A and B are called **mutually exclusive** if $A \cap B = \varnothing$.

Events A_1, A_2, A_3, ..., are said to be **mutually exclusive** if they are pairwise mutually exclusive. That is, if $A_i \cap A_j = \varnothing$ whenever $i \neq j$.

Definition 1.1.6: Exhaustive

Events $A_1, A_2, A_3, \dots$ are said to be **exhaustive** if $A_1 \cup A_2 \cup A_3 \cup \cdots = S$.

1.2 Definition of Probability

Definition 1.2.1: Probability

For a given experiment, S denotes the sample space and $A_1, A_2, A_3, \dots$ represent possible events. A set function that associates a real value $\mathbb{P}(A)$ with each event A is called a **probability set function**, and $\mathbb{P}(A)$ is called the **probability** of A, if the following properties are satisfied:

- $\mathbb{P}(A) \geq 0$ for every A
- $\mathbb{P}(S) = 1$
- $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, if $A_1, A_2, A_3, \dots$ are pairwise mutually exclusive events.

Definition 1.2.2

If an object is chosen from a finite collection of distinct objects in such a manner that each object has the same probability of being chosen, then we say that the object was chosen **at random**.

1.3 Some Properties of Probability

Theorem 1.3.1

If A is an event and A' is its complement, then

$$\mathbb{P}(A) = 1 - \mathbb{P}(A'). \tag{1.3.1}$$

Theorem 1.3.2

For any event A, $\mathbb{P}(A) \leq 1$.

Theorem 1.3.3

For any two events A and B,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \tag{1.3.2}$$

Theorem 1.3.4

For any three events A, B, and C,

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) = {} & \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ & - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ & + \mathbb{P}(A \cap B \cap C). \end{aligned} \tag{1.3.3}$$

Theorem 1.3.5

If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Theorem 1.3.6: Boole's Inequality

If $A_1, A_2, \ldots$ is a sequence of events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{1.3.4}$$

Theorem 1.3.7: Bonferroni's Inequality

If $A_1, A_2, \ldots, A_K$ are events, then

$$\mathbb{P}\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} \mathbb{P}(A_i'). \tag{1.3.5}$$

1.4 Conditional Probability

Definition 1.4.1: Conditional Probability

The **conditional probability** of an event A, given the event B, is defined by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{1.4.1}$$

if $\mathbb{P}(B) \neq 0$.

Theorem 1.4.1: Multiplication Theorem

For any events A and B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B) = \mathbb{P}(A)\mathbb{P}(B|A). \tag{1.4.2}$$

Theorem 1.4.2: Total Probability

If $B_1, B_2, \dots, B_k$ is a collection of mutually exclusive and exhaustive events, then for any event A,

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(B_i)\mathbb{P}(A|B_i). \tag{1.4.3}$$

Theorem 1.4.3: Bayes' Rule

If $B_1, B_2, \dots, B_k$ is a collection of mutually exclusive and exhaustive events, then for any event A,

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\mathbb{P}(A|B_j)}{\sum\limits_{i=1}^{k} \mathbb{P}(B_i)\mathbb{P}(A|B_i)}. \tag{1.4.4}$$

Definition 1.4.2: Independent & Dependent

Two events A and B are called **independent events** if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{1.4.5}$$

Otherwise, A and B are called **dependent events**.

Theorem 1.4.4

If A and B are events such that $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then A and B are independent if and only if either of the following holds:

$$\mathbb{P}(A|B) = \mathbb{P}(A), \quad \mathbb{P}(B|A) = \mathbb{P}(B).$$

Theorem 1.4.5

Two events A and B are independent if and if the following pairs of events are also independent:

1. A and B'.
2. A' and B.
3. A' and B'.

Definition 1.4.3: Mutually Independent

The k events $A_1, A_2, \dots, A_k$ are said to be **independent** or **mutually independent** if for every $j = 2, 3, \ldots, k$ and every subset of distinct indices $i_1, i_2, \ldots, i_j$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_j}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_j}). \tag{1.4.6}$$

1.5 Counting Techniques

If the i-th of r successive operations can be performed in n_i ways, then the total number of ways to carry out all r operations is the product

$$\prod_{i=1}^{r} n_i = n_1\, n_2 \cdots n_r. \tag{1.5.1}$$

Theorem 1.5.1

If there are N possible outcomes of each of r trials of an experiment, then there are N^r possible outcomes in the sample space.

Definition 1.5.1: Indistinguishable & Distinguishable

Two elements are called **indistinguishable** if a new result or arrangement will not be obtained when they are interchanged. Otherwise, the two elements are called **distinguishable**.

An ordered arrangement of a set of objects is known as a **permutation**.

Theorem 1.5.2

The number of permutations of n distinguishable objects is $n!$.

Theorem 1.5.3

The number of permutations of n distinct objects taken r at a time is

$$_nP_r = \frac{n!}{(n-r)!}. \tag{1.5.2}$$

If the order of the objects is not important, then one may simply be interested in the number of **combinations** that are possible when selecting r objects from n distinct objects. The symbol $\binom{n}{r}$ usually is used to denote this number.

Theorem 1.5.4

The number of combinations of n distinct objects taken r at a time is

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}. \tag{1.5.3}$$

Theorem 1.5.5

The number of permutations of n objects of which r_1 are of one kind, r_2 of a second kind, $\dots$, r_k of a k-th kind is

$$\frac{n!}{r_1!\, r_2! \cdots r_k!}. \tag{1.5.4}$$

where $\sum_{i=1}^{k} r_i = n$.

Theorem 1.5.6

The number of ways of partitioning a set of n objects into k cells with r_1 objects in the first cell, r_2 in the second cell, and so forth is

$$\frac{n!}{r_1!\, r_2! \cdots r_k!}. \tag{1.5.5}$$

where $\sum_{i=1}^{k} r_i = n$.

Chapter 2

Random Variables and Their Distributions

2.1 Introduction

Definition 2.1.1: Random Variables

A random variable, say X, is a function defined over a sample space, S, that associates a real number, $X(e) = x$, with each possible outcome e in S.

2.2 Cumulative Distribution Function

Definition 2.2.1: CDF

The **cumulative distribution function** (CDF) of a random variable X is defined for any real x by

$$F(x) = \mathbb{P}(X \leq x). \tag{2.2.1}$$

Theorem 2.2.1

A function $F(x)$ is a CDF for some random variable X if and only if it satisfies the following properties:

- $\lim_{x \to -\infty} F(x) = 0;$
- $\lim_{x \to \infty} F(x) = 1;$
- $\lim_{h \to 0^+} F(x+h) = F(x);$
- $a < b$ implies $F(a) \leq F(b)$.

A probability distribution for a random variable X is of **mixed type** if the CDF has the form

$$F(x) = aF_1(x) + (1-a)F_2(x) \tag{2.2.2}$$

where $F_1(x)$ and $F_2(x)$ are CDFs of discrete and continuous type, respectively, and $0 < a < 1$.

2.3 Random Variables

Discrete Random Variables

Definition 2.3.1: Discrete Random Variable & PMF (Discrete PDF)

If the set of all possible values of a random variable, X is a countable set, $x_1, x_2, \ldots, x_n$ or $x_1, x_2, \ldots$, then X is called a **discrete random variable**. The function

$$f(x) = \mathbb{P}(X = x), \quad x = x_1,\ x_2,\ \ldots \tag{2.3.1}$$

that assigns the probability to each possible value x will be called the **discrete probability density function** (discrete PDF) or **probability mass function** (PMF).

Theorem 2.3.1

A function $f(x)$ is a discrete PDF if and only if it satisfies both of the following properties for at most a countably infinite set of reals $x_1, x_2, \ldots$:

- $f(x_i) \geq 0$ for all x_i;
- $\sum\limits_{\text{all } x_i} f(x_i) = 1.$

Definition 2.3.2

If X is a discrete random variable with discrete PDF $f(x)$, then the **expected value** of X is defined by

$$\mathbb{E}(X) = \sum_x x f(x). \tag{2.3.2}$$

Other common notations for $\mathbb{E}(X)$ include μ or μ_X, and the terms **mean** or **expection** of X is often used.

Continuous Random Variables

Definition 2.3.3: Continuous Random Variable & PDF

A random variable X is called a **continuous random variable** if there is a function $f(x)$, called the **probability density function** (PDF) of X, such that the CDF can be represented as

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t. \tag{2.3.3}$$

Theorem 2.3.2

A function $f(x)$ is a PDF if and only if it satisfies both of the following properties x_1, x_2, . . . :

- $f(x) \geq 0$ for all real x;
- $\displaystyle\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$

Definition 2.3.4

If X is a continuous random variable with PDF $f(x)$, then the **expected value** of X is defined by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \tag{2.3.4}$$

if the integral is absolutely convergent. Otherwise we say that $\mathbb{E}(X)$ does not exist.

Other common notations for $\mathbb{E}(X)$ include μ or μ_X, and the terms **mean** or **expection** of X is often used.

Other Definitions

Definition 2.3.5: Mode

If the PDF has a unique maximum at $x = m_0$, say $f(m_0) = \max f(x)$, then m_0 is called the **mode** of X.

Definition 2.3.6: Symmetric

A distribution with PDF $f(x)$ is said to be **symmetric** about c if $f(c - x) = f(c + x)$ for all x.

2.4 Some Properties of Expected Values

Theorem 2.4.1

If X is a random variable with PDF $f(x)$ and $u(x)$ is a real-valued function whose domain includes the possible values of X, then

$$\mathbb{E}[u(X)] = \sum_x u(x)f(x), \qquad \text{if } X \text{ is discrete;} \tag{2.4.1}$$

$$\mathbb{E}[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)\,\mathrm{d}x, \qquad \text{if } X \text{ is continuous.} \tag{2.4.2}$$

Theorem 2.4.2

If X is a random variable with PDF $f(x)$, a and b are constants, and $g(x)$ and $h(x)$ are real-valued functions whose domains include the possible values of X, then

$$\mathbb{E}[ag(X) + bh(X)] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)]. \tag{2.4.3}$$

Definition 2.4.1: Variance

The **variance** of a random variable X is given by

$$\operatorname{Var}(X) = \mathbb{E}[(X - \mu)^2]. \tag{2.4.4}$$

Other common notations for $\operatorname{Var}(X)$ are σ^2 or σ_X^2, and a related quantity, called **standard deviation** of X, is the positive square root of the variance, $\sigma = \sigma_X = \sqrt{\operatorname{Var}(X)}$.

Theorem 2.4.3

If X is a random variable, then

$$\operatorname{Var}(X) = \mathbb{E}(X^2) - \mu^2. \tag{2.4.5}$$

Theorem 2.4.4

If X is a random variable and a and b are constants, then

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X). \tag{2.4.6}$$

Theorem 2.4.5

If the distribution of X is symmetric about the mean $\mu = \mathbb{E}(X)$, then the third moment about μ is zero, $\mu_3 = 0$.

Bounds on Probability

Theorem 2.4.6

If X is a random variable and $u(x)$ is a non-negative real-valued function, then for any positive constant $c > 0$,

$$\mathbb{P}[u(X) \geq c] \leq \frac{\mathbb{E}[u(X)]}{c}. \tag{2.4.7}$$

Proof If $A = \{x | u(x) \geq c\}$, then for a continuous random variable,

$$\begin{aligned}\mathbb{E}[u(X)] &= \int_{-\infty}^{\infty} u(x) f(x)\,\mathrm{d}x \\ &= \int_A u(x) f(x)\,\mathrm{d}x + \int_{A^c} u(x) f(x)\,\mathrm{d}x \\ &\geq \int_A u(x) f(x)\,\mathrm{d}x \\ &\geq \int_A c f(x)\,\mathrm{d}x \\ &= c\,\mathbb{P}(x \in A) \\ &= c\,\mathbb{P}[u(X) \geq c].\end{aligned}$$

A similar proof holds for discrete variables. □

Theorem 2.4.7: Markov's Inequality

If X is a random variable, then for any positive constant $c > 0$,

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}(|X|^r)}{c^r}. \tag{2.4.8}$$

Proof Let $u(x) = |x|^r$ for $r > 0$ in Theorem 2.4.6. □

Theorem 2.4.8: Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2, then for any $k > 0$,

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}. \tag{2.4.9}$$

Proof Let $u(x) = (X - \mu)^2$ and $c = k^2\sigma^2$ in Theorem 2.4.6. □

Lemma 2.4.1: Hoeffding's Lemma

Let X be any real-valued random variable with expected value $\mathbb{E}(X) = 0$ and such that $a \leq X \leq b$ almost surely. Then, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left(e^{\lambda X}\right) \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right). \tag{2.4.10}$$

Proof Since $e^{\lambda x}$ is a convex function of x, we have

$$e^{\lambda X} \leq \frac{X-a}{b-a} e^{\lambda b} + \frac{b-X}{b-a} e^{\lambda a}.$$

Take expectation of both sides and use the that $\mathbb{E}(X) = 0$ to get

$$\mathbb{E}\left(e^{\lambda X}\right) \leq -\frac{a}{b-a} e^{\lambda b} + \frac{b}{b-a} e^{\lambda a} = e^{g(u)},$$

where $u = \lambda(b-a)$, $g(u) = -pu + \log(1-p+pe^u)$ and $p = -a/(b-a)$.

Note that $g(0) = g'(0) = 0$. Also, $g''(u) \leq 1/4$ for all $u > 0$. By Taylor's theorem, there is $\xi \in (0,u)$ such that

$$g(u) = g(0) + ug'(0) + \frac{1}{2}u^2 g''(\xi) = \frac{1}{2}u^2 g''(\xi) \leq \frac{u^2}{8}.$$

Hence,

$$\mathbb{E}\left(e^{\lambda X}\right) \leq e^{g(u)} \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

□

Theorem 2.4.9: Hoeffding's Inequality

Let $X_1, \dots, X_n$ be independent observation such that $\mathbb{E}(X_i) = 0$ and $a_i \leq X_i \leq b_i$. Let $t > 0$, then

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{2.4.11}$$

Proof For any $s > 0$, we have, from Markov's inequality, that

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) = \mathbb{P}\left(s\sum_{i=1}^n X_i \geq st\right) &= \mathbb{P}\left(e^{s\sum_{i=1}^n X_i} \geq e^{st}\right) \\ &\leq e^{-st}\mathbb{E}\left(e^{s\sum_{i=1}^n X_i}\right) = e^{-st}\prod_{i=1}^n \mathbb{E}\left(e^{sX_i}\right).\end{aligned}$$

Then, from Hoeffding's lemma,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq e^{-st}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8} = \exp\left(-st + \frac{1}{8}s^2\sum_{i=1}^n (b_i-a_i)^2\right).$$

To get the best possible upper bound, we find the minimum of

$$\begin{cases} g : \mathbb{R}_+ \mapsto \mathbb{R}, \\ g(s) = -st + \frac{1}{8}s^2\sum_{i=1}^n (b_i - a_i)^2. \end{cases}$$

Note that g is a quadratic function and achieves its minimum at $s = 4t/\sum_{i=1}^n (b_i - a_i)^2$. Thus we get the result. □

Theorem 2.4.10: Cauchy-Schwarz Inequality

If X and Y are two random variables, The Cauchy-Schwarz inequality states that:

$$[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2) \cdot \mathbb{E}(Y^2). \tag{2.4.12}$$

Theorem 2.4.11: Jensen's Inequality

Let X a random variable such that $\mathbb{E}[|X|] < \infty$. Jensen's inequality states that

- if $f : \mathbb{R} \mapsto \mathbb{R}$ is a convex function, then $f(\mathbb{E}(X)) \leq \mathbb{E}[f(X)]$;
- if $f : \mathbb{R} \mapsto \mathbb{R}$ is a concave function, then $\mathbb{E}[f(X)] \leq f(\mathbb{E}(X))$.

Theorem 2.4.12

Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$. If $\sigma^2 = 0$, then $\mathbb{P}(X = \mu) = 1$.

Approximate Mean and Variance

If a function of a random variable, say $H(X)$, can be expanded in a Taylor series, then the function $H(x)$ has a Taylor approximation about μ:

$$H(x) \doteq H(\mu) + H'(\mu)(x-\mu) + \tfrac{1}{2}H''(\mu)(x-\mu)^2. \tag{2.4.13}$$

which suggests the approximation

$$\mathbb{E}[H(x)] \doteq H(\mu) + \tfrac{1}{2}H''(\mu)\sigma^2, \tag{2.4.14}$$

and, using the first two terms,

$$\mathrm{Var}[H(x)] \doteq [H'(\mu)]^2\sigma^2. \tag{2.4.15}$$

2.5 Moment Generating Functions

Definition 2.5.1: Moments

The k-th **moment about the origin** of a random variable X is

$$\mu_k' = \mathbb{E}(X^k), \tag{2.5.1}$$

and the k-th **moment about the mean** is

$$\mu_k = \mathbb{E}[(X-\mu)^k]. \tag{2.5.2}$$

Theorem 2.5.1

The general equation for converting the n-th order moment about the origin to the moment about the mean is

$$\mu_n = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} \mu_j' \mu^{n-j}, \tag{2.5.3}$$

where μ is the mean of the distribution, and the moment about the origin is given by

$$\mu'_m = \sum_{j=0}^{m} \binom{m}{j} \mu_j \mu^{m-j}. \tag{2.5.4}$$

Definition 2.5.2: MGF

If X is a random variable, then the expected value

$$M_X(t) = \mathbb{E}(\mathrm{e}^{tX}) \tag{2.5.5}$$

is called the **moment generating function** (MGF) of X if this expected value exists for all values of t in some interval of the form $-h < t < h$ for some $h > 0$.

Theorem 2.5.2

If the MGF of X exists, then

$$\mu'_n = \mathbb{E}(X^n) = M_X^{(n)}(0) \quad \text{for all } n = 1, 2, \ldots \tag{2.5.6}$$

and

$$M_X(t) = 1 + \sum_{n=1}^{\infty} \frac{\mu'_n t^n}{n!}. \tag{2.5.7}$$

Proof

$$e^{tX} = 1 + \sum_{n=1}^{\infty} \frac{X^n t^n}{n!} \quad \Rightarrow \quad \mathbb{E}(\mathrm{e}^{tX}) = 1 + \sum_{n=1}^{\infty} \frac{\mu'^n t^n}{n!}.$$ □

Theorem 2.5.3

If $Y = aX + b$, then $M_Y(t) = \mathrm{e}^{bt} M_X(at)$.

Theorem 2.5.4: Uniqueness

If X_1 and X_2 have respective CDFs $F_1(x)$ and $F_2(x)$, and MGFs $M_1(t)$ and $M_2(t)$, then $F_1(x) = F_2(x)$ for all real x if and only $M_1(t) = M_2(t)$ for all t in some interval $-h < t < h$ for some $h > 0$.

Cumulant

Definition 2.5.3: CGF

The cumulant generating function (CGF) of a random variable X is defined as the natural logarithm of the MGF:

$$K_X(t) = \ln \mathbb{E}\left(e^{tX}\right) = \ln M_X(t). \tag{2.5.8}$$

Definition 2.5.4: Cumulants

If the CGF of X exists, then the cumulants is defined as

$$\kappa_n = K_X^{(n)}(0) \quad \text{for all } n = 1, 2, \ldots \tag{2.5.9}$$

Theorem 2.5.5

If the CGF of X exists, then

$$K_X(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} = \mu t + \sigma^2 \frac{t^2}{2} + \ldots. \tag{2.5.10}$$

Factorial Moments

Definition 2.5.5

The r-th **factorial moment** of X is

$$\mathbb{E}[X(X-1)\cdots(X-r+1)], \tag{2.5.11}$$

and the **factorial moment generating function** (FMGF) of X is

$$G_X(t) = \mathbb{E}(t^X) \tag{2.5.12}$$

if this expection exists for all t in some interval of the form $1 - h < t < 1 + h$.

Also note that the FMGF sometimes is called the **probability generating function**.

Theorem 2.5.6

If X has a FMGF, $G_X(t)$, then

$$G_X^{(r)}(1) = \mathbb{E}[X(X-1)\cdots(X-r+1)]. \tag{2.5.13}$$

2.6 Skewness and Kurtosis

The skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.

Definition 2.6.1: Skewness

The skewness of a random variable X is the third standardized moment γ_1, defined as:

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E}(X^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \tag{2.6.1}$$

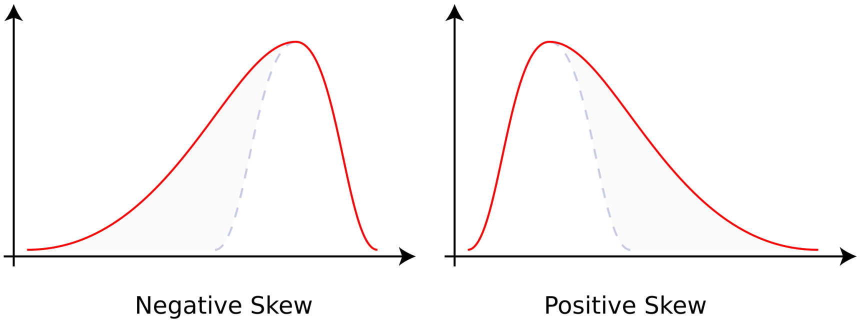

Figure 2.1: Skewness

The kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable.

Definition 2.6.2: Kurtosis

The kurtosis is the fourth standardized moment, defined as

$$\gamma_2 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}. \tag{2.6.2}$$

Theorem 2.6.1

The kurtosis is bounded below by the squared skewness plus 1:

$$\gamma_2 \geq \gamma_1^2 + 1 \quad \Rightarrow \quad \frac{\mu_4}{\sigma^4} \geq \left(\frac{\mu_3}{\sigma^3}\right)^2 + 1. \tag{2.6.3}$$

Chapter 3

Joint Distributions

3.1 Joint Discrete Distributions

Definition 3.1.1

The **joint probability density function** (joint PDF) of the k-dimensional discrete random variable $X = (X_1, X_2, \dots, X_k)$ is defined to be

$$f(x_1, x_2, \dots, x_k) = \mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_k = x_k) \tag{3.1.1}$$

for all possible values $x = (x_1, x_2, \dots, x_k)$ of X.

Theorem 3.1.1

A function $f(x_1, x_2, \dots, x_k)$ is the joint PDF for some vector-valued random variable $X = (X_1, X_2, \dots, X_k)$ if and only if the following properties are satisfied:

- $f(x_1, x_2, \dots, x_k) \geq 0$ for all possible values $x = (x_1, x_2, \dots, x_k)$ of X;
- $\sum\limits_{x_1} \cdots \sum\limits_{x_k} f(x_1, x_2, \dots, x_k) = 1$.

Definition 3.1.2

If the pair (X_1, X_2) of discrete random variables has the joint PDF $f(x_1, x_2)$, then the **marginal PDFs** of X_1 and X_2 are

$$f_1(x_1) = \sum_{x_2} f(x_1, x_2), \quad f_2(x_2) = \sum_{x_1} f(x_1, x_2). \tag{3.1.2}$$

Definition 3.1.3: Joint CDF

The **joint cumulatice distribution function** of the k random variables X_1, X_2, $\dots$, X_k is the function defined by

$$F(x_1,\dots,x_k) = \mathbb{P}(X_1 \le x_1,\dots,X_k \le x_k). \tag{3.1.3}$$

Theorem 3.1.2

A function $F(x_1,x_2)$ is a bivariate CDF if and only if

- $\lim\limits_{x_1\to-\infty} F(x_1,x_2) = F(-\infty,x_2) = 0$ for all x_2;
- $\lim\limits_{x_2\to-\infty} F(x_1,x_2) = F(x_1,-\infty) = 0$ for all x_1;
- $\lim\limits_{\substack{x_1\to\infty\\x_2\to\infty}} F(x_1,x_2) = F(\infty,\infty) = 1$;
- $\mathbb{P}(a < X_1 \le b, c < X_2 \le d) = F(b,d) - F(b,c) - F(a,d) + F(a,c) \ge 0$;
- $\lim\limits_{h\to0^+} F(x_1+h,x_2) = \lim\limits_{h\to0^+} F(x_1,x_2+h) = F(x_1,x_2)$ for all x_1 and x_2.

3.2 Joint Continuous Distributions

Definition 3.2.1

A k-dimensional discrete random variable $X = (X_1,X_2,\dots,X_k)$ is said to be **continuous** if there is a function $f(x_1,x_2,\dots,x_k)$, called the **joint probability density function** (joint PDF), of X, such that the joint CDF can be written as

$$F(x_1,x_2,\dots,x_k) = \int_{-\infty}^{x_k}\cdots\int_{-\infty}^{x_1} f(t_1,\dots,t_k)\ \mathrm{d}t_1\cdots\mathrm{d}t_k. \tag{3.2.1}$$

for all $x = (x_1,x_2,\dots,x_k)$.

Theorem 3.2.1

Any function $f(x_1,\dots,x_k)$ is the joint PDF of a k-dimensional random variable $X = (X_1,\dots,X_k)$ if and only if the following properties are satisfied:

- $f(x_1,\dots,x_k) \ge 0$ for all $x = (x_1,\dots,x_k)$;
- $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\dots,x_k)\,\mathrm{d}x_1\cdots\mathrm{d}x_k = 1$.

Definition 3.2.2

If the pair (X_1,X_2) of continuous random variables has the joint PDF $f(x_1,x_2)$, then the

marginal PDFs of X_1 and X_2 are

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_2, \quad f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2)\,\mathrm{d}x_1. \tag{3.2.2}$$

Definition 3.2.3

If $X = (X_1, \ldots, X_k)$ is a k-dimensional random variable with joint CDF $F(x_1, \ldots, x_k)$, then the **marginal CDF** of X_j is

$$F_j(x_j) = \lim_{\substack{x_i \to \infty \\ \text{all } i \neq j}} F(x_1, \ldots, x_j, \ldots, x_k). \tag{3.2.3}$$

3.3 Independent Random Variables

Definition 3.3.1: Independent Random Variables

Random variables $X_1, \ldots, X_k$ are said to be **independent** if for every $a_i < b_i$,

$$\mathbb{P}[a_1 \leq X_1 \leq b_1, \ldots, a_k \leq X_k \leq b_k] = \prod_{i=1}^{k} \mathbb{P}[a_i \leq X_i \leq b_i]. \tag{3.3.1}$$

Theorem 3.3.1

Random variables $X_1, \ldots, X_k$ are independent if and only if the following properties holds:

$$F(x_1, \ldots, x_k) = F_1(x_1) \cdots F_k(x_k), \tag{3.3.2}$$
$$f(x_1, \ldots, x_k) = f_1(x_1) \cdots f_k(x_k), \tag{3.3.3}$$

where $F_i(x_i)$ and $f_i(x_i)$ are the marginal CDF and PDF of X_i, respectively.

Theorem 3.3.2

Two random variables X_1 and X_2 with joint PDF $f(x_1, x_2)$ are independent if and only if:

1. the “support set”, $(x_1, x_2) | f(x_1, x_2) > 0$, is a Cartesian product, $A \times B$;
2. the joint PDF can be factored into the product of functions of x_1 and x_2, $f(x_1, x_2) = g(x_1)\,h(x_2)$.

3.4 Conditional Distributions

Definition 3.4.1: Conditional PDF

If X_1 and X_2 are continuous (or discrete) random variables with joint PDF $f(x_1, x_2)$, then the **conditional PDF** of X_2 given $X_1 = x_1$ is defined to be

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}. \tag{3.4.1}$$

for values x_1 such that $f_1(x_1) > 0$, and zero otherwise.

Theorem 3.4.1

If X_1 and X_2 are random variables with joint PDF $f(x_1, x_2)$ and marginal PDFs $f_1(x_1)$ and $f_2(x_2)$, then

$$f(x_1, x_2) = f_1(x_1)f(x_2|x_1) = f_2(x_2)f(x_1|x_2). \tag{3.4.2}$$

And if X_1 and X_2 are independent, then

$$f(x_2|x_1) = f_2(x_2), \quad f(x_1|x_2) = f_1(x_1). \tag{3.4.3}$$

3.5 Random Samples

Definition 3.5.1: Random Sample

The set of random variables $X_1, \dots, X_n$ is said to be a **random sample** of size n from a population with density function $f(x)$ if the joint PDF has the form

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2)\cdots f(x_n). \tag{3.5.1}$$

Empirical Distributions

We now take a set of data $x_1, x_2, \dots, x_n$ from a random sample of size n from $f(x)$, and let $y_1 < y_2 < \cdots < y_n$ be the **ordered** values of the data. Then the **empirical CDF** based on this data can be represented as

$$\hat{F}_n(x) = \begin{cases} 0, & x < y_1, \\ i/n, & y_i \le x < y_{i+1}, \\ 1, & y_n \le x. \end{cases} \tag{3.5.2}$$

Histograms

It usually is easier to study the distribution of probability in terms of the PDF, $f(x)$, rather than the CDF. This leads us to consider a different type of empirical distribution, known as a **histogram**.

Chapter 4

Properties of Random Variables

4.1 Probabilities of Expected Values

Definition 4.1.1

If $X = (X_1, \dots, X_k)$ has a joint PDF $f(x_1, \dots, x_k)$, and if $Y = u(X_1, \dots, X_k)$ is a function of X, then

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}_X[u(X_1, \dots, X_k)] \\ &= \begin{cases} \sum_{x_1} \cdots \sum_{x_k} u(x_1, \dots, x_k) f(x_1, \dots, x_k), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, \dots, x_k) f(x_1, \dots, x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_k, & \text{if } X \text{ is continuous.} \end{cases}\end{aligned}$$

Theorem 4.1.1

If $X_1, X_2, \dots, X_k$ are jointly distributed random variables and $a_1, a_2, \dots, a_k$ are constants, then

$$\mathbb{E}\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i \mathbb{E}(X_i). \tag{4.1.1}$$

Theorem 4.1.2

If X and Y are independent random variables and $g(x)$ and $h(y)$ are functions, then

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\ \mathbb{E}[h(Y)]. \tag{4.1.2}$$

Definition 4.1.2: Covariance

The **covariance** of a pair of random variables X and Y is defined by

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]. \tag{4.1.3}$$

Another common notation for covariance is σ_{XY}.

Theorem 4.1.3

If X and Y are random variables and a and b are constants then

$$\begin{aligned} \operatorname{Cov}(Y, X) &= \operatorname{Cov}(X, Y), && (4.1.4)\\ \operatorname{Cov}(aX, bY) &= ab \operatorname{Cov}(X, Y), && (4.1.5)\\ \operatorname{Cov}(X + a, Y + b) &= \operatorname{Cov}(X, Y), && (4.1.6)\\ \operatorname{Cov}(X, aX + b) &= a \operatorname{Var}(X). && (4.1.7) \end{aligned}$$

Theorem 4.1.4

If $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_m$ are jointly random distributed variables, and if $a_1, \ldots, a_k$ and $b_1, \ldots, b_n$ are constants, then

$$\operatorname{Cov}\left(\sum_{i=1}^{k} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{m} a_i b_j \operatorname{Cov}(X_i, Y_j). \tag{4.1.8}$$

Theorem 4.1.5

If X and Y are random variables, then

$$\operatorname{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{4.1.9}$$

and $\operatorname{Cov}(X, Y) = 0$ whenever X and Y are independent.

Theorem 4.1.6

If X and Y are random variables, then

$$[\operatorname{Cov}(X, Y)]^2 \le \operatorname{Var}(X) \cdot \operatorname{Var}(Y). \tag{4.1.10}$$

Proof By Cauchy-Schwarz inequalities, we have

$$\begin{aligned} [\operatorname{Cov}(X, Y)]^2 &= \{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]\}^2\\ &\le \mathbb{E}[(X - \mu_X)^2] \cdot \mathbb{E}[(Y - \mu_Y)^2]\\ &= \operatorname{Var}(X) \cdot \operatorname{Var}(Y). \end{aligned}$$

□

Theorem 4.1.7

If $X_1, X_2, \ldots, X_k$ are jointly distributed random variables and $a_1, a_2, \ldots, a_k$ are constants, then

$$\operatorname{Var}\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i^2 \operatorname{Var}(X_i) + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} a_i a_j \operatorname{Cov}(X_i, X_j). \tag{4.1.11}$$

4.2 Approximate Mean and Variance

Consider a pair of random variables (X, Y) with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 and covariance σ_{12}; further suppose that the function $H(x, y)$ has partial derivatives in an open rectangle containing (μ_1, μ_2). Using Taylor approximations, we obtain the following approximate formulas for the mean and variance of $H(X, Y)$:

$$\mathbb{E}[H(X,Y)] \doteq H(\mu_1, \mu_2) + \frac{\partial^2 H}{\partial x^2}\sigma_1^2 + \frac{\partial^2 H}{\partial y^2}\sigma_2^2, \tag{4.2.1}$$

$$\operatorname{Var}[H(X,Y)] \doteq \left(\frac{\partial H}{\partial x}\right)^2 \sigma_1^2 + \left(\frac{\partial H}{\partial y}\right)^2 \sigma_2^2 + 2\frac{\partial H}{\partial x}\frac{\partial H}{\partial y}\sigma_{12}, \tag{4.2.2}$$

where the partial derivatives are evaluated at the means (μ_1, μ_2).

4.3 Correlation

Definition 4.3.1: Correlation Coefficient

If X and Y are random variables with variances σ_X^2 and σ_Y^2 and covariance $\sigma_{XY} = \operatorname{Cov}(X, Y)$, then the **correlation coefficient** of X and Y is

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{4.3.1}$$

Theorem 4.3.1

If ρ is the correlation coefficient of X and Y, then $-1 \le \rho \le 1$. And $\rho = \pm 1$ if and only if $Y = aX + b$ with probability 1 for some $a \neq 0$ and b.

Proof Let $W = Y/\sigma_Y - \rho X/\sigma_X$, so that

$$\begin{aligned}\operatorname{Var}(W) &= \left(\frac{1}{\sigma_Y}\right)\sigma_Y^2 + \left(\frac{\rho}{\sigma_X}\right)\sigma_X^2 - 2\rho\frac{\sigma_{XY}}{\sigma_X\sigma_Y} \\ &= 1 + \rho^2 - 2\rho^2 = 1 - \rho^2 \ge 0.\end{aligned}$$

Since $\rho = \pm 1$, we have $\operatorname{Var}(W) = 0$, so $\mathbb{P}(W = \mu_W) = 1$, which means with probability 1,

$$W = \frac{Y}{\sigma_Y} - \rho\frac{X}{\sigma_X} = \frac{\mu_Y}{\sigma_Y} - \rho\frac{\mu_X}{\sigma_X} \quad \Leftrightarrow \quad Y = aX + b,$$

where $a = \rho\sigma_Y/\sigma_X$, $b = \mu_Y - \rho\mu_X\sigma_Y/\sigma_X$. On the other hand, if $Y = aX + b$, then $\sigma_Y^2 = a^2\sigma_X^2$ and $\sigma_{XY} = a\sigma_X^2$, in which case $\rho = a/|a|$, so that $\rho = 1$ if $a > 0$ and $\rho = -1$ if $a < 0$. □

4.4 Conditional Expectation

Definition 4.4.1: Conditional Expectation

If X and Y are jointly distributed random variables, then the **conditional expectation** of Y given $X = x$ is given by

$$\mathbb{E}(Y|x) = \begin{cases} \sum_y y f(y|x), & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} y f(y|x)\,\mathrm{d}y, & \text{if } Y \text{ is continuous.} \end{cases} \tag{4.4.1}$$

Theorem 4.4.1: Law of Total Expectation

If X and Y are jointly distributed random variables, then

$$\mathbb{E}_X[\mathbb{E}(Y|X)] = \mathbb{E}(Y). \tag{4.4.2}$$

Proof Consider the continuous case:

$$\begin{aligned} \mathbb{E}_X[\mathbb{E}(Y|X)] &= \int_{-\infty}^{\infty} \mathbb{E}(Y|X) f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f(y|x)\,\mathrm{d}y\right] f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y \\ &= \mathbb{E}(Y). \end{aligned}$$

□

Theorem 4.4.2

If X and Y are independent random variables, then $\mathbb{E}(Y|x) = \mathbb{E}(Y)$ and $\mathbb{E}(X|y) = \mathbb{E}(X)$.

Definition 4.4.2: Conditional Variance

The **conditional variance** of Y given $X = x$ is given by

$$\mathrm{Var}(Y|x) = \mathbb{E}\left[(Y - \mathbb{E}(Y|x)^2|x\right] = \mathbb{E}[Y^2|x] - (\mathbb{E}(Y|x))^2. \tag{4.4.3}$$

Theorem 4.4.3: Law of Total Variance

If X and Y are jointly distributed random variables, then

$$\mathrm{Var}(Y) = \mathbb{E}_X[\mathrm{Var}(Y|X)] + \mathrm{Var}_X[\mathbb{E}(Y|X)]. \tag{4.4.4}$$

Proof

$$\begin{aligned}\mathbb{E}_X[\mathrm{Var}(Y|X)] &= \mathbb{E}_X[\mathbb{E}[Y^2|X] - (\mathbb{E}(Y|X))^2] \\ &= \mathbb{E}[Y^2] - \mathbb{E}_X[(\mathbb{E}(Y|X))^2] \\ &= \mathbb{E}[Y^2] - [\mathbb{E}(Y)]^2 - \{\mathbb{E}_X[(\mathbb{E}(Y|X))^2] - [\mathbb{E}(Y)]^2\} \\ &= \mathrm{Var}(Y) - \mathrm{Var}_X[\mathbb{E}(Y|X)].\end{aligned}$$

□

Theorem 4.4.4

If X and Y are jointly distributed random variables and $h(x,y)$ is a function, then

$$\mathbb{E}[h(X,Y)] = \mathbb{E}_X\Big[\mathbb{E}[h(X,Y)|X]\Big]. \tag{4.4.5}$$

Theorem 4.4.5

If X and Y are jointly distributed random variables, and $g(x)$ is a function, then

$$\mathbb{E}[g(X)Y|x] = g(x)\mathbb{E}(Y|x). \tag{4.4.6}$$

Theorem 4.4.6

If $\mathbb{E}(Y|x)$ is a linear function of x, then

$$\mathbb{E}(Y|x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \tag{4.4.7}$$

$$\mathbb{E}_X[\mathrm{Var}(Y|X)] = \sigma_Y^2(1-\rho^2). \tag{4.4.8}$$

Proof If $\mathbb{E}(Y|x) = ax+b$, then

$$\mu_Y = \mathbb{E}[Y] = \mathbb{E}_X[\mathbb{E}(Y|X)] = \mathbb{E}_X(aX+b) = a\mu_X + b,$$

and

$$\begin{aligned}\sigma_{XY} &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[(X-\mu_X)Y] \\ &= \mathbb{E}_X[\mathbb{E}[(X-\mu_X)Y|X]] = \mathbb{E}_X[(X-\mu_X)\mathbb{E}(Y|X)] \\ &= \mathbb{E}_X[(X-\mu_X)(aX+b)] = a\sigma_X^2.\end{aligned}$$

Thus,

$$a = \frac{\sigma_{XY}}{\sigma_X^2} = \rho\frac{\sigma_Y}{\sigma_X}, \quad b = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X.$$

$$\begin{aligned}\mathbb{E}_X[\mathrm{Var}(Y|X)] &= \mathrm{Var}(Y) - \mathrm{Var}_X[\mathbb{E}(Y|X)] \\ &= \mathrm{Var}(Y) - \mathrm{Var}_X\left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)\right] \\ &= \sigma_Y^2 - \left(\rho\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2 = \sigma_Y^2(1-\rho^2).\end{aligned}$$

□

Bivariate Normal Distribution

A pair of continuous random variables X and Y is said to have a **bivariate normal distribution** if it has a joint PDF of the form

$$\begin{aligned} f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\Bigg\{ -\frac{1}{2(1-\rho^2)} \\ \left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\Bigg\} \end{aligned} \tag{4.4.9}$$

for $x, y \in (-\infty, \infty)$. A special notation for this is

$$(X,Y) \sim \mathsf{BVN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho). \tag{4.4.10}$$

Theorem 4.4.7

If $(X,Y) \sim \mathsf{BVN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and ρ is the correlation coefficient of X and Y.

Theorem 4.4.8

If $(X,Y) \sim \mathsf{BVN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then

1. conditional on $X = x$,

$$Y|x \sim \mathcal{N}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2)\right);$$

2. conditional on $Y = y$,

$$X|y \sim \mathcal{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2), \sigma_1^2(1-\rho^2)\right).$$

4.5 Joint Moment Generating Functions

Definition 4.5.1: Joint MGF

The joint MGF of $X = (X_1, \ldots, X_k)$, if it exists, is defined to be

$$M_X(t) = \mathbb{E}\left[\exp\left(\sum_{i=1}^{k} t_i X_i\right)\right], \tag{4.5.1}$$

where $\boldsymbol{t} = (t_1, \ldots, t_k)$ and $-h < t_i < h$ for some $h > 0$.

Theorem 4.5.1

If $M_{X,Y}(t_1, t_2)$ exists, then the random variables X and Y are independent if and only if $M_{X,Y}(t_1, t_2) = M_X(t_1)M_Y(t_2)$.

Chapter 5

Functions of Random Variables

5.1 The CDF Technique

We will assume that a random variable X has CDF $F_x(x)$, and that some function of X is of interest, say $Y = u(X)$. The idea behind the CDF technique is to express the CDF of Y in terms of the distribution of X. Specifically, for each real y, we can define a set $A_y = \{x_i | u(x) \leq y\}$. It follows that $[Y \leq y]$ and $[X \in A_y]$ are equivalent events, and consequently

$$F_Y(y) = \mathbb{P}[u(X) \leq y] \tag{5.1.1}$$

which also can be expressed as $\mathbb{P}[X \in A_y]$. This probability can be expressed as the integral of the PDF $F_X(x)$, over the set A_y if X is continuous, or the summation off $F_X(x)$ over X in A_y if X is discrete.

For example, it often is possible to express $[u(X) \leq y]$ in terms of an equivalent event $[x_1 \leq X \leq x_2]$, where one or both of the limits x_1 and x_2 depend on y.

In the continuous case,

$$F_Y(y) = \int_{x_1}^{x_2} f_X(x) \, \mathrm{d}x = F_X(x_2) - F_X(x_1) \tag{5.1.2}$$

and, of course, the PDF is $f_Y(y) = (\mathrm{d}/\mathrm{d}y) F_Y(y)$.

Theorem 5.1.1

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ be a k-dimensional vector continuous random variables, with joint PDF $f(x_1, x_2, \ldots, x_k)$. If $Y = u(X)$ is a function of X, then

$$F_Y(y) = \mathbb{P}[u(X) \leq y] = \underset{A_y}{\int \cdots \int} f(x_1, x_2, \ldots, x_k) \, \mathrm{d}x_1 \cdots \mathrm{d}x_k, \tag{5.1.3}$$

where $A_y = \{\boldsymbol{x} | u(\boldsymbol{x}) \leq y\}$.

5.2 Transformation Methods

One-To-One Transformations

Theorem 5.2.1

Suppose that X is a discrete random variable with PDF $F_X(x)$ and that $Y = u(X)$ defines a one-to-one transformation. In other words, the equation $y = u(x)$ can be solved uniquely, say $x = w(y)$. Then the PDF of Y is

$$f_Y(y) = f_X(w(y)), \quad y \in B, \tag{5.2.1}$$

where $B = \{y|f_Y(y) > 0\}$.

Theorem 5.2.2

Suppose that X is a continuous random variable with PDF $f_X(x)$, and assume that Y = u(X) defines a one-to-one transformation from $A = \{x|f_X(x) > 0\}$ on to $B = \{y|f_Y(y) > 0\}$ with inverse transformation $x = w(y)$. If the derivative $(\mathrm{d}/\mathrm{d}y)w(y)$ is continuous and nonzero on B, then the PDF of Y is

$$f_Y(y) = f_X(w(y)) \left|\frac{\mathrm{d}}{\mathrm{d}y}w(y)\right|, \quad y \in B. \tag{5.2.2}$$

Transformations That Are Not One-To-One

Suppose that the function $u(x)$ is not one-to-one over $A = \{x|f_X(x) > 0\}$. Although this means that no unique solution to the equation $y = u(x)$ exists, it usually is possible to partition A into disjoint subsets A_1, A_2, $\ldots$, such that $u(x)$ is one-to-one over each A_j. Then, for each y in the range of $u(x)$, the equation $y = u(x)$ has a unique solution $x_j = w_j(y)$ over the set A_j. Then the PDF of Y can be given by

$$f_Y(y) = \sum_j f_X(w_j(y)). \tag{5.2.3}$$

Joint Transformations

Theorem 5.2.3

If $\boldsymbol{X}$ is a vector of discrete random variables with joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x})$ and $\boldsymbol{Y} = \boldsymbol{u}(\boldsymbol{X})$ defines a one-to-one transformation, then the joint PDF of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_k) = f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k) \tag{5.2.4}$$

where $x_1, x_2, \ldots, x_k$ are the solutions of $\boldsymbol{y} = \boldsymbol{u}(\boldsymbol{x})$, and consequently depend on y_1, y_2, $\ldots$, y_k.

If the transformation is not one-to-one, and if a partition exists, say A_1, A_2, $\ldots$, such that the equation $\boldsymbol{y} = \boldsymbol{u}(\boldsymbol{x})$ has a unique solution $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{kj})$ over A_j, then the PDF of Y is

$$f_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_k) = \sum_j f_{\boldsymbol{X}}(x_{1j}, x_{2j}, \ldots, x_{kj}). \tag{5.2.5}$$

Theorem 5.2.4

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ is a vector of continuous rardom variables with joint PDF $f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k)$ on A, and $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)$ is defined by the one-to-one transformation

$$Y_i = u_i(X_1, X_2, \ldots, X_k), \quad i = 1, 2, \ldots, k. \tag{5.2.6}$$

If the Jacobian is continuous and nonzero over the range of the transformation, then the joint PDF of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(y_1, y_2, \ldots, y_k) = f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_k)\,|J| \tag{5.2.7}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ is the solution of $\boldsymbol{y} = \boldsymbol{u}(\boldsymbol{x})$ and

$$J = \left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right| = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \cdots & \dfrac{\partial x_1}{\partial y_k} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial x_k}{\partial y_1} & \cdots & \dfrac{\partial x_k}{\partial y_k} \end{vmatrix}. \tag{5.2.8}$$

Example 5.2.1

Find the PDF of $Z = X_1/X_2$, where X_1 and X_2 are continuous with joint PDF $f(x_1, x_2)$.

Solution Let $Y_1 = X_2$, $Y_2 = Z = X_1/X_2$, then $x_1 = y_1 y_2$, $x_2 = y_1$, and the joint PDF of Y is

$$f_{\boldsymbol{Y}}(y_1, y_2) = f(y_1 y_2, y_1) \begin{vmatrix} y_2 & y_1 \\ 1 & 0 \end{vmatrix} = |y_1|\, f(y_1 y_2, y_1).$$

Therefore, the marginal PDF of $Z = Y_2$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_{\boldsymbol{Y}}(y_1, z)\, \mathrm{d}y_1 = \int_{-\infty}^{\infty} |y|\, f(yz, y)\, \mathrm{d}y.$$

5.3 The Quantile Transformation

Definition 5.3.1: Quantile Function

For any CDF F, the **quantile function** (also called **percent-point function** or **inverse cumulative distribution function**) is defined by

$$Q(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\} \quad \text{for } 0 < u < 1. \tag{5.3.1}$$

Note that the infimum function can be replaced by the minimum function, since the distribution function is right-continuous and weakly monotonically increasing. Figure 5.1 gives illustrations of CDF and QF.

If the CDF function $F(x) = \mathbb{P}(X \leq x)$ is continuous and strictly monotonically increasing, then the inequalities can be replaced by equalities, and we have:

$$Q = F^{-1}. \tag{5.3.2}$$

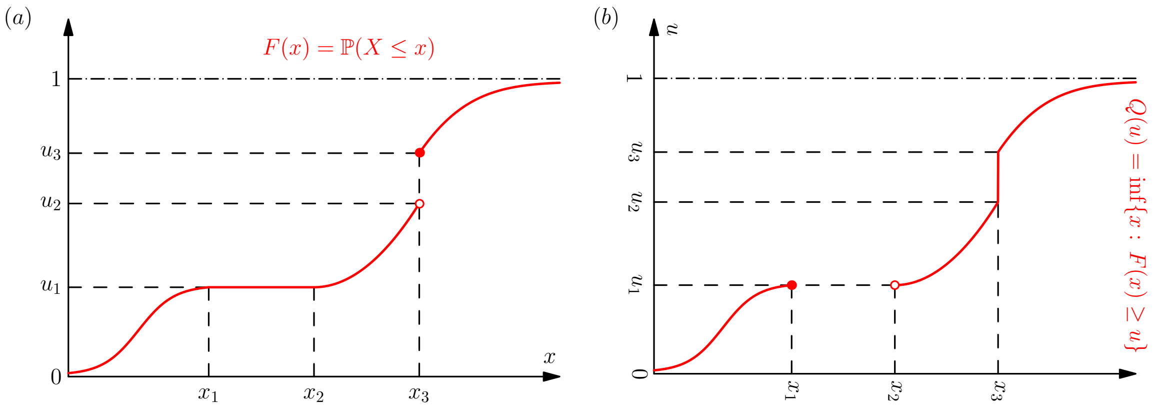

Figure 5.1: Illustrations of CDF and QF.

Theorem 5.3.1

Let F be a CDF and Q be a QF, then

1. $u \le F(x) \Leftrightarrow Q(u) \le x$ for all real x.
2. If $U \sim \mathsf{UNIF}(0,1)$, then $X = Q(U) \sim F$.

Proof 1. For a given x_0:

- $\Rightarrow$: $u \le F(x_0)$ implies $Q(u) \le x_0$ since $x_0 \in \{x \in \mathbb{R} : F(x) \ge u\}$.
- $\Leftarrow$: If $Q(u) = \inf\{x \in \mathbb{R} : F(x) \ge u\} \le x_0$, then, for all $\varepsilon > 0$, $x_0 + \varepsilon \in \{x \in \mathbb{R} : F(x) \ge u\}$ and $F(x_0 + \varepsilon) \ge u$. Thus, $F(x_0) \ge u$ since $F(\cdot)$ is right-continuous.

2. The CDF of $X = Q(U)$ is given by $F_X(x) = \mathbb{P}[Q(U) \le x]$. With the result of the first part, $\mathbb{P}[Q(U) \le x] = \mathbb{P}[U \le F(x)] = F(x)$. □

Lemma 5.3.1

Let X have CDF F. Then for all real x, $\mathbb{P}[F(X) \le F(x)] = F(x)$.

Proof Decompose the event:

$$\{F(X) \le F(x)\} = \Big[\{F(X) \le F(x)\} \cap \{X \le x\}\Big] \bigcup \Big[\{F(X) \le F(x)\} \cap \{X > x\}\Big].$$

Since $\{X \le x\} \subset \{F(X) \le F(x)\}$ and $\{X > x\} \cap \{F(X) < F(x)\} = \varnothing$, it follows that

$$\{F(X) \le F(x)\} = \{X \le x\} \bigcup \Big[\{X > x\} \cap \{F(X) = F(x)\}\Big].$$

Taking probabilities, the result follows because the last event in brackets has probability 0 (since it implies that X lies in the interior of an interval of constancy of F). □

Theorem 5.3.2: Probability Integral Transformation

If X is continuous with CDF F, then $U = F(X) \sim \mathsf{UNIF}(0,1)$.

Proof Let $u \in (0,1)$. Since F is continuous, there exists a real x such that $F(x) = u$:

1. If F is strictly increasing, then $Q = F^{-1}$. There exists only one x such that $F(x) = u$, i.e. $x = Q(u) = F^{-1}(u)$.
2. If F is flat and suppose $F(x) = u$ for all $x \in [x_1, x_2)$, then $Q(u) = x_1$.

Then by Lemma 5.3.1, $\mathbb{P}(U \leq u) = \mathbb{P}[F(X) \leq F(x)] = F(x) = u$, which implies that $U \sim \mathsf{UNIF}(0,1)$. □

5.4 Sums of Random Variables

Convolution Formula

If one is interested only in the PDF of a sum $S = X_1 + X_2$, where X_1 and X_2 are continuous with joint PDF $f(x_1, x_2)$, then a general formula can be given by

$$f_S(s) = \int_{-\infty}^{\infty} f(t, s-t)\,\mathrm{d}t = \int_{-\infty}^{\infty} f(s-t, t)\,\mathrm{d}t. \tag{5.4.1}$$

If X_1 and X_2 are independent, then

$$f_S(s) = \int_{-\infty}^{\infty} f_1(t) f_2(s-t)\,\mathrm{d}t = \int_{-\infty}^{\infty} f_1(s-t) f_2(t)\,\mathrm{d}t. \tag{5.4.2}$$

Proof The CDF of S is

$$\begin{aligned}
F_S(s) &= \mathbb{P}[S \leq s] = \mathbb{P}[X_1 + X_2 \leq s] \\
&= \iint\limits_{x_1 + x_2 \leq s} f(x_1, x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{s - x_1} f(x_1, x_2)\,\mathrm{d}x_2 \right] \mathrm{d}x_1 \\
\xrightarrow{x_2 = u - x_1} &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{s} f(x_1, u - x_1)\,\mathrm{d}u \right] \mathrm{d}x_1 = \int_{-\infty}^{s} \left[\int_{-\infty}^{\infty} f(t, u-t)\,\mathrm{d}t \right] \mathrm{d}u,
\end{aligned}$$

Then, the PDF of S is

$$f_S(s) = \frac{\mathrm{d}F_S(s)}{\mathrm{d}s} = \int_{-\infty}^{\infty} f(t, s-t)\,\mathrm{d}t.$$

□

Moment Generating Function Method

Theorem 5.4.1

If $X_1, \ldots, X_n$ are independent random variables with MGFs $M_{X_i}(t)$, then the MGF of $Y = \sum_{i=1}^{n} X_i$ is

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t). \tag{5.4.3}$$

5.5 Ordered Statistics

Let $X_1, X_2, \dots, X_n$ be a random sample of size n, and the joint PDF of the associated n independent random variables is given by

$$f(x_1, x_2, \dots, x_n) = f(x_1) f(x_2) \cdots f(x_n). \tag{5.5.1}$$

We will consider a transformation that orders the observations $x_1, x_2, \dots, x_n$. For example, let $y_i = u_i(x_1, x_2, \dots, x_n)$ represents the i-th smallest of $x_1, x_2, \dots, x_n$. When this transformation is applied to a random sample $X_1, X_2, \dots, X_n$, we will obtain a set of ordered random variables, called the **order statistics** and denoted by either $X_{1:n}, X_{2:n}, \dots, X_{n:n}$ or $Y_1, Y_2, \dots, Y_n$.

Theorem 5.5.1

If $X_1, X_2, \dots, X_n$ is a random sample from a population with continuous PDF $f(x)$, then the joint PDF of the order statistics $Y_1, Y_2, \dots, Y_n$, is

$$g(y_1, y_2, \dots, y_n) = n! f(y_1) f(y_2) \cdots f(y_n) \tag{5.5.2}$$

if $y_1 < y_2 < \cdots < y_n$, and zero otherwise.

Proof The sample space of ordered random sample:

$$B = \{(y_1, y_2, \dots, y_n) | y_1 < y_2 < \cdots < y_n\}$$

can be partitioned into the following $n!$ disjoint sets:

$$\begin{aligned} A_1 &= \{(x_1, x_2, \dots, x_n) | x_1 < x_2 < \cdots < x_n\}, \\ A_2 &= \{(x_2, x_1, \dots, x_n) | x_2 < x_1 < \cdots < x_n\}, \\ &\cdots \end{aligned}$$

In transforming to the ordered random sample, we have the one-to-one transformation

$$\begin{aligned} Y_1 = X_1,\ Y_2 = X_2,\ \dots,\ Y_n = X_n \quad &\text{with } J_1 = 1 \text{ on } A_1, \\ Y_1 = X_2,\ Y_2 = X_1,\ \dots,\ Y_n = X_n \quad &\text{with } J_2 = -1 \text{ on } A_2, \\ &\cdots \end{aligned}$$

Notice that in each case $|J_i| = 1$. Furthermore, for each region, the joint PDF is the product of factors $f(y_i)$ multiplied in some order, but can be written regardless of the order. If we sum over all $n!$ subsets, then the joint PDF of $Y_1, Y_2, \dots, Y_n$ is (5.5.2). □

Theorem 5.5.2

Suppose that $X_1, X_2, \dots, X_n$ denotes a random sample of size n from a continuous PDF, $f(x)$, where $f(x) > 0$ for $a < x < b$. Then the PDF of the k-th order statistic Y_i, is given by

$$\begin{aligned} g_k(y_k) &= \frac{n!}{(k-1)!1!(n-k)!}[\mathbb{P}(X \leq y_k)]^{k-1} f(y_k)[\mathbb{P}(X \geq y_k)]^{n-k} \\ &= \frac{n!}{(k-1)!(n-k)!}[F(y_k)]^{k-1} f(y_k)[1-F(y_k)]^{n-k} \end{aligned} \tag{5.5.3}$$

if $a < y_k < b$, and zero otherwise.

The PDF of a pair of order statistics Y_i and Y_j where $i < j$ is given by

$$\begin{aligned} g_{ij}(y_i, y_j) &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[\mathbb{P}(X \leq y_i)]^{i-1} f(y_i) \\ &\quad \times [\mathbb{P}(y_i \leq X \leq y_j)]^{j-i-1} f(y_j)[\mathbb{P}(X \geq y_j)]^{n-j} \\ &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(y_i)]^{i-1} f(y_i) \\ &\quad \times [F(y_j) - F(y_i)]^{j-i-1} f(y_j)[1-F(y_j)]^{n-j} \end{aligned} \tag{5.5.4}$$

if $a < y_i < y_j < b$, and zero otherwise.

The smallest and largest order statistics are of special importance, as are certain functions of order statistics known as the sample median and range. If n is odd, then the **sample median** is the middle observation, Y_K where $k = (n+1)/2$; if n is even, then it is considered to be any value between the two middle observations Y_k and Y_{k+1} where $k = n/2$, although it is often taken to be their average. The **sample range** is the difference of the smallest from the largest, $R = Y_n - Y_1$. For continuous random variables, the PDFs of the minimum and maximum, Y_1 and Y_n, which are special cases of equation (5.5.3) and (5.5.4) are

$$g_1(y_1) = nf(y_1)[1-F(y_1)]^{n-1}, \qquad a < y_1 < b, \tag{5.5.5}$$

$$g_n(y_n) = n[F(y_n)]^{n-1} f(y_n), \qquad a < y_n < b, \tag{5.5.6}$$

$$g_{1n}(y_1, y_n) = n(n-1)f(y_1)[F(y_n) - F(y_1)]^{n-2} f(y_n), \qquad a < y_1 < y_n < b. \tag{5.5.7}$$

Theorem 5.5.3

For a random sample of size n from a discrete or continuous CDF, $F(x)$, the marginal CDF of the k-th order statistic is given by

$$G_k(y_k) = \sum_{j=k}^{n} \binom{n}{j} [F(y_k)]^j [1-F(y_k)]^{n-j}. \tag{5.5.8}$$

Censored Sampling

In certain types of problems such as life-testing experiments, the ordered observations may occur naturally. In such cases a great savings in time and cost may be realized by terminating the experiment after only the first r ordered observations have occurred, rather than waiting for all n failures to occur. The usually is referred to as **Type II Censored Sampling**.

Theorem 5.5.4: Type II Censored Sampling

The joint marginal density function of the first r order statistics from a random sample of size n from a continuous PDF, $f(x)$, is given by

$$g(y_1, y_2, \dots, y_n) = \frac{n!}{(n-r)!}[1 - F(y_r)]^{n-r} \prod_{i=1}^{r} f(y_i) \tag{5.5.9}$$

if $-\infty < y_1 < \cdots < y_r < \infty$, and zero otherwise.

Proof

$$g(y_1, y_2, \dots, y_n) = \binom{n}{r} r! \prod_{i=1}^{r} f(y_i) \cdot [1 - F(y_r)]^{n-r}. \qquad \square$$

In Type II censored sampling the number of observations, r, is fixed but the length of experiment, Y_r, is a random variable. If one terminates the experiment after a fixed time t_0, this procedure is referred to as **Type I Censored Sampling**. In this case the number of observations, R, is a random variable. The probability that a failure occurs before time t_0 for any given trial is $p = F(t_0)$, so for a random sample of size n the random variable R follows a binomial distribution:

$$R \sim \mathsf{BIN}(n, F(t_0)). \tag{5.5.10}$$

Type I censored sampling is related to the concept of truncated sampling and truncated distributions. Consider a random variable X with pdf $f(x)$ and CDF $F(x)$. If it is given that a random variable from this distribution has a value less than t_0, then the CDF of X given $X \le t_0$ is referred to as the *truncated distribution* of X, truncated on the right at t_0, and is given by

$$F(x|x \le t_0) = \frac{\mathbb{P}[X \le x,\ X \le t_0]}{Pr[X \le t_0]} = \frac{F(x)}{F(t_0)}, \quad 0 < x < t_0, \tag{5.5.11}$$

and

$$f(x|x \le t_0) = f(x)/F(t_0), \quad 0 < x < t_0. \tag{5.5.12}$$

Distributions truncated on the left are defined similarly.

Theorem 5.5.5: Type I Censored Sampling

If $Y_1, \dots, Y_r$ denote the observed values of a random sample of size n from $f(x)$ that is Type I censored on the right at t_0, then the joint PDF of $Y_1, \dots, Y_R$ is given by

$$f_{Y_1,\dots,Y_R}(y_1, \dots, y_r) = \frac{n!}{(n-r)!}[1 - F(t_0)]^{n-r} \prod_{i=1}^{r} f(y_i) \tag{5.5.13}$$

if $y_1 < \cdots < y_r < t_0$ and $r = 1, 2, \dots, n$, and

$$\mathbb{P}(R = 0) = [1 - F(t_0)]^n. \tag{5.5.14}$$

Proof

$$f_{Y_1,\dots,Y_R}(y_1, \dots, y_r) = g(y_1, \dots, y_r|r) b(r; n, F(t_0)),$$

where

$$g(y_1, \ldots, y_r|r) = \frac{r!}{[F(t_0)^r]} \prod_{i=1}^{r} f(y_i), \quad y_1 < \cdots < y_r < t_0$$

and

$$b(r; n, F(t_0)) = \binom{n}{r} [F(t_0)]^r [1 - F[t_0]]^{n-r}. \qquad \square$$

Chapter 6

Limiting Distributions

6.1 Sequence of Random Variables

Definition 6.1.1: Converge In Distribution

If $Y_n \sim G_n(y)$ for each $n = 1, 2, \dots,$ and if for some CDF $G(y)$,

$$\lim_{n\to\infty} G_n(y) = G(y) \tag{6.1.1}$$

for all values y at which $G(y)$ is continuous, then the sequence $Y_1, Y_2, \dots$ is said to **converge in distribution** to $Y \sim G(y)$, denoted by $Y_n \xrightarrow{d} Y$. The distribution corresponding to the CDF $G(y)$ is called the **limiting distribution** of Y_n.

Definition 6.1.2: Degenerate Distribution

The function $G(y)$ is the CDF of a **degenerate distribution** at the value $y = c$ if

$$G(y) = \begin{cases} 0, & y < c, \\ 1, & y \geq c. \end{cases} \tag{6.1.2}$$

In other words, $G(y)$ is the CDF of a discrete distribution that assigns probability one at the value $y = c$ and zero otherwise.

Definition 6.1.3: Converge in Probability

The sequence of random variables Y_n is said to be **convergence in probability** to Y, written $Y_n \xrightarrow{p} Y$, if

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| < \varepsilon) = 1, \quad \forall\, \varepsilon > 0. \tag{6.1.3}$$

Definition 6.1.4: Almost Sure Convergence

The sequence of random variables Y_n is said to be **almost sure convergence** (or **conver-**

gence with probability 1) to Y, written $Y_n \xrightarrow{a.s.} Y$, if

$$\mathbb{P}\left(\lim_{n\to\infty} Y_n = Y\right) = 1. \tag{6.1.4}$$

Definition 6.1.5: Convergence in r-th mean

The sequence of random variables Y_n is said to be **convergence in r-th mean** (or **in the L^r norm**) to Y, written $Y_n \xrightarrow{L^r} Y$, if

$$\lim_{n\to\infty} \mathbb{E}(|Y_n - Y|^r) = 0. \tag{6.1.5}$$

Provided the probability space is complete, the relationship between the definitions of convergence:

$$\begin{tikzcd}
Y_n \xrightarrow{L^s} Y \arrow[r, Rightarrow, "s > r \geq 1"'] & Y_n \xrightarrow{L^r} Y \arrow[d, Rightarrow, "r > 0"'] & \\
Y_n \xrightarrow{a.s} Y \arrow[r, Rightarrow] & Y_n \xrightarrow{p} Y \arrow[r, Rightarrow] & Y_n \xrightarrow{d} Y
\end{tikzcd}$$

Theorem 6.1.1

For a sequence of random variables Y_n, then

1. $Y_n \xrightarrow{a.s.} Y$ implies $Y_n \xrightarrow{p} Y$.
2. $Y_n \xrightarrow{p} Y$ implies $Y_n \xrightarrow{d} Y$.
3. $Y_n \xrightarrow{p} c$ for a constant c if and only if $Y_n \xrightarrow{d} c$.

Theorem 6.1.2

If $Y_n \xrightarrow{p} c$, then for any function $g(y)$ that is continuous at c, $g(Y_n) \xrightarrow{p} g(c)$.

Proof Because $g(y)$ is continuous at c, it follows that for every $\varepsilon > 0$ a $\delta > 0$ exists such that $|y - c| < \delta$ implies $|g(y) - g(c)| < \varepsilon$. This, in turn, implies that

$$\mathbb{P}[|g(Y_n) - g(c)| < \varepsilon] \geq \mathbb{P}(|Y_n - c| < \delta)$$

because $\mathbb{P}(B) \geq p(A)$ whenever $A \subset B$. But because $Y_n \xrightarrow{p} c$, it follows for every $\delta > 0$ that

$$\lim_{n\to\infty} \mathbb{P}[|g(Y_n) - g(c)| < \varepsilon] \geq \lim_{n\to\infty} \mathbb{P}(|Y_n - c| < \delta) = 1.$$

The left-hand limit cannot exceed 1, so it must equal 1, and $g(Y_n) \xrightarrow{p} g(c)$. □

Theorem 6.1.3: Continuous Mapping

Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(Y \in C) = 1$.

1. If $Y_n \xrightarrow{d} Y$, then $g(Y_n) \xrightarrow{d} g(Y)$;
2. If $Y_n \xrightarrow{p} Y$, then $g(Y_n) \xrightarrow{p} g(Y)$;
3. If $Y_n \xrightarrow{a.s.} Y$, then $g(Y_n) \xrightarrow{a.s.} g(Y)$.

Theorem 6.1.4

If X_n and Y_n are two sequences of random variables such that $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{p} d$, then

1. $aX_n + bY_n \xrightarrow{p} ac + bd$.
2. $X_nY_n \xrightarrow{p} cd$.
3. $X_n/c \xrightarrow{p} 1$, for $c \neq 0$.
4. $1/X_n \xrightarrow{p} 1/c$ if $\mathbb{P}(X_n \neq 0) = 1$ for all $n, c \neq 0$.
5. $\sqrt{X_n} \xrightarrow{p} \sqrt{c}$ if $\mathbb{P}(X_n \geq 0) = 1$ for al n.

Theorem 6.1.5: Slutsky's Theorem

If X_n and Y_n are two sequences of random variables such that $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{d} Y$, then

1. $X_n + Y_n \xrightarrow{d} c + Y$.
2. $X_nY_n \xrightarrow{d} cY$.
3. $Y_n/X_n \xrightarrow{d} Y/c$, for $c \neq 0$.

Theorem 6.1.6

Let $Y_1, Y_2, \ldots$ be a sequence of random variables with respective CDFs $G_1(y), G_2(y), \ldots$ and MGFs $M_1(y)$, $M_2(y)$, $\ldots$ If $M(t)$ is the MGF of a CDF $G(y)$, and if $\lim_{n\to\infty} M_n(t) = M(t)$ for all t in an open interval containing zero, $-h < t < h$, then $\lim_{n\to\infty} G_n(t) = G(t)$ for all continuity points of $G(y)$.

6.2 Law of Large Numbers

Theorem 6.2.1: Weak Law of Large Numbers (Khintchine's law)

If $\{X_1, X_2, \ldots\}$ is a sequence of i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ for all i, then the sample average converges in probability towards the expected value:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu, \quad \text{as } n \to \infty. \tag{6.2.1}$$

That is,

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| < \varepsilon\right) = 1, \quad \forall\, \varepsilon > 0. \tag{6.2.2}$$

Proof **Method I:** Uses the assumption of finite variance $\text{Var}(X_i) = \sigma^2 < \infty$.
The independence of the random variables implies

$$\mathbb{E}(\bar{X}_n) = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\text{Var}(X_1) + \cdots + \text{Var}(X_n)}{n} = \frac{\sigma^2}{n},$$

Using Chebyshev's inequality (2.4.9) on $\bar{X}_n$ results in

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \quad \Rightarrow \quad \mathbb{P}(|\bar{X}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

As n approaches infinity, the expression approaches 1.

Method II:

According to the Taylor series formula, the MGF of X can be written as

$$M_X(t) = 1 + \mu t + \frac{M_X''(\xi)t^2}{2},$$

where ξ is between 0 and t. Therefore,

$$M_{\bar{X}_n}(t) = M_{\sum X_i}\left(\frac{t}{n}\right) = \left[M_X\left(\frac{t}{n}\right)\right]^n = \left[1 + \mu\frac{t}{n} + \frac{M_X''(\xi_n)t^2}{2n^2}\right]^n$$

where ξ_n is between 0 and t/n. As $n \to \infty$, $\xi_n \to 0$, $M_X''(\xi_n) \to \sigma^2$ and $M_{\bar{X}_n}(t) \to e^{\mu t}$, which is the MGF of degenerate distribution at μ. □

Theorem 6.2.2: Strong Law of Large Numbers

If $\{X_1, X_2, \ldots\}$ is a sequence of i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ for all i, then the sample average converges almost surely to the expected value:

$$\bar{X}_n \xrightarrow{a.s.} \mu, \quad \text{as } n \to \infty. \tag{6.2.3}$$

That is,

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1. \tag{6.2.4}$$

6.3 The Central Limit Theorem

Theorem 6.3.1: Central Limit Theorem (Lindeberg-Lévy)

If $\{X_1, X_2, \dots\}$ is a sequence of i.i.d. random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$ for all i, then the limiting distribution of

$$Z_n = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} \tag{6.3.1}$$

is the standard normal, $Z_n \xrightarrow{d} Z \sim \mathcal{N}(0,1)$ as $n \to \infty$.

Proof According to the Taylor series formula, the MGF of X can be written as

$$M_{X-\mu}(t) = 1 + \frac{M''_{X-\mu}(\xi)t^2}{2} = 1 + \frac{\sigma^2 t^2}{2} + \frac{(M''_X(\xi_n) - \sigma^2)t^2}{2},$$

where ξ is between 0 and t. Therefore,

$$M_{Z_n}(t) = M_{\sum(X_i-\mu)}\left(\frac{t}{\sqrt{n}\sigma}\right) = \left[M_{X-\mu}\left(\frac{t}{\sqrt{n}\sigma}\right)\right]^n = \left[1 + \frac{\sigma^2 t^2}{2n\sigma^2} + \frac{(M''_X(\xi_n) - \sigma^2)t^2}{2n\sigma^2}\right]^n,$$

where ξ_n is between 0 and $t/(\sqrt{n}\sigma)$. As $n \to \infty$, $\xi_n \to 0$, $M''_X(\xi_n) \to \sigma^2$ and $M_{Z_n}(t) \to \mathrm{e}^{t^2/2}$, which is the MGF of standard normal distribution. □

Theorem 6.3.2: Central Limit Theorem (Lyapunov)

Suppose $\{X_1, X_2, \dots\}$ is a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2. Define

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2. \tag{6.3.2}$$

If for some $\delta > 0$, Lyapunov's condition

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left(|X_i - \mu_i|^{2+\delta}\right) = 0 \tag{6.3.3}$$

is satisfied, then as $n \to \infty$:

$$Z_n = \frac{1}{s_n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{d} Z \sim \mathcal{N}(0,1). \tag{6.3.4}$$

6.4 Asymptotic Normal Distributions

Definition 6.4.1: Asymptotic Normal

If $Y_1, Y_2, \dots$ is a sequence of random variables and m and c are constants such that

$$Z_n = \frac{Y_n - m}{c/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1) \tag{6.4.1}$$

as $n \to \infty$, then Y_n is said to have an **asymptotic normal distribution** with **asymptotic mean** m and **asymptotic variance** c^2/n, or approximately, $Y_n \sim \mathcal{N}(m, c^2/n)$ as $n \to \infty$.

Theorem 6.4.1

If $Z_n = \sqrt{n}(Y_n - m)/c \xrightarrow{d} Z \sim \mathcal{N}(0,1)$, then $Y_n \xrightarrow{p} m$.

Theorem 6.4.2: Delta Method

If $\sqrt{n}(Y_n - m)/c \xrightarrow{d} Z \sim \mathcal{N}(0,1)$ and if $g'(m) \neq 0$, then

$$\frac{\sqrt{n}[g(Y_n) - g(m)]}{|cg'(m)|} \xrightarrow{d} Z \sim \mathcal{N}(0,1), \tag{6.4.2}$$

or approximately, $g(Y_n) \sim \mathcal{N}(g(m), [cg'(m)]^2/n)$ as $n \to \infty$.

Proof **Method I:** Define

$$u(y) = \begin{cases} \dfrac{g(y) - g(m)}{y - m} - g'(m), & y \neq m, \\ 0, & y = m. \end{cases}$$

It follows that $u(y)$ is continuous at m with $u(m) = 0$, and thus $g'(m) + u(Y_n) \xrightarrow{p} g'(m)$. Further more

$$\frac{\sqrt{n}[g(Y_n) - g(m)]}{cg'(m)} = \frac{\sqrt{n}(Y_n - m)}{c} \frac{g'(m) + u(Y_n)}{g'(m)}.$$

From Theorem 6.1.4, we have $[g'(m) + u(Y_n)]/g'(m) \xrightarrow{p} 1$, and the result follows from Slutsky's Theorem 6.1.5.

Method II: According to mean value theorem, there exists ξ between Y_n and m such that

$$g(Y_n) = g(m) + (Y_n - m)g'(\xi).$$

Since $\xi \xrightarrow{p} m$ as $Y_n \xrightarrow{p} m$, we have $g'(\xi) \xrightarrow{p} g'(m)$ if g' is continuous at m according to Theorem 6.1.2. Therefore,

$$\frac{\sqrt{n}[g(Y_n) - g(m)]}{cg'(m)} = \frac{\sqrt{n}(Y_n - m)}{c} \frac{g'(\xi)}{g'(m)}.$$

From Theorem 6.1.4, we have $g'(\xi)/g'(m) \xrightarrow{p} 1$, and the result follows from Slutsky's Theorem 6.1.5. □

Theorem 6.4.3

If $\sqrt{n}(Y_n - m)/c \xrightarrow{d} Z \sim \mathcal{N}(0,1)$ and if $g'(m) = 0$, then

$$\frac{2n[g(Y_n) - g(m)]}{c^2 g''(m)} \xrightarrow{d} \chi^2(1), \tag{6.4.3}$$

provided $g''(m)$ exist and is not zero.

Proof According to mean value theorem, there exists ξ between Y_n and m such that

$$g(Y_n) = g(m) + \frac{1}{2}(Y_n - m)^2 g''(\xi).$$

Since $\xi \xrightarrow{p} m$ as $Y_n \xrightarrow{p} m$, we have $g''(\xi) \xrightarrow{p} g'(m)$ if g'' is continuous at m according to Theorem 6.1.2. Therefore,

$$\frac{2n[g(Y_n) - g(m)]}{c^2 g''(m)} = \frac{n(Y_n - m)^2}{c^2} \frac{g''(\xi)}{g''(m)}.$$

From Theorem 6.1.4, we have $g''(\xi)/g''(m) \xrightarrow{p} 1$, and the result follows from Slutsky's Theorem 6.1.5 and Theorem 7.2.2.

Example 6.4.1

Let S_n^2 denote the sample variance from a random sample of size n from $\mathcal{N}(\mu, \sigma^2)$. We know that

$$V_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1),$$

and from Theorem 7.2.5,

$$\frac{\sqrt{n-1}(S_n^2 - \sigma^2)}{\sigma^2\sqrt{2}} = \frac{V_n - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

or approximately,

$$S_n^2 \sim \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n-1}\right).$$

If $Y_n = S_n^2$, and $g(y) = \sqrt{y}$, then $g'(\sigma) = 1/(2\sqrt{\sigma})$, and approximately

$$S_n \sim \mathcal{N}\left(\sigma, \frac{\sigma^2}{2(n-1)}\right).$$

Asymptotic Distributions of Central Order Statistics

Theorem 6.4.4

Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with a PDF $f(x)$ that is continuous and nonzero at the p-th percentile, x_p, for $0 < p < 1$. If $k/n \to p$ (with $k - np$ bounded), then the sequence of k-th order statistics, $X_{k:n}$, is asymptotically normal with mean x_p, and variance c^2/n, where

$$c^2 = \frac{p(1-p)}{[f(x_p)]^2}. \tag{6.4.4}$$

Chapter 7

Statistics and Sampling Definitions

7.1 Statistics

Consider a set of observable random variables $X_1, \dots, X_n$. For example, suppose the variables are a random sample of size n from a population.

Definition 7.1.1: Statistic

A function of observable random variables, $T = \mathscr{T}(X_1, \dots, X_n)$, which does not depend on any unknown parameters, is called a **statistic**.

Definition 7.1.2: Sample Moments

1. The k-th **sample moment about the origin** of random variables $X_1, \dots, X_n$ is

$$M_k' = \frac{1}{n} \sum_{i=1}^{n} X_i^k. \tag{7.1.1}$$

2. The k-th **sample moment about the mean** is

$$M_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^k, \tag{7.1.2}$$

   where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the sample mean. Note $M_1 \equiv 0$.

3. The k-th **sample central moment** (only defined when $\mu_i = \mathbb{E}(X_i)$ is known) is

$$C_k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_i)^k. \tag{7.1.3}$$

Theorem 7.1.1

Let $X_1, \ldots, X_n$ denotes a random sample, and define the following power sum

$$S_k = \sum_{i=1}^{n} X_i^k, \quad k = 0, 1, \ldots, \tag{7.1.4}$$

then

$$M_k = \frac{1}{n} \sum_{j=0}^{k} \binom{k}{j} \frac{(-1)^j}{n^j} S_1^j S_{k-j}, \quad k = 2, 3, \ldots. \tag{7.1.5}$$

Proof Using binomial theorem, we have

$$M_k = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \frac{S_1}{n} \right)^k = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{k} \binom{k}{j} X_i^{k-j} \frac{(-1)^j S_1^j}{n^j} = \frac{1}{n} \sum_{j=0}^{k} \binom{k}{j} \frac{(-1)^j}{n^j} S_1^j S_{k-j}. \qquad \square$$

Theorem 7.1.2

Let $X_1, \ldots, X_n$ denotes an IID (independently and identically distributed) random sample, then

$$\mathbb{E}(M_2) = \frac{n-1}{n} \mu_2, \tag{7.1.6a}$$

$$\mathbb{E}(M_3) = \frac{(n-1)(n-2)}{n^2} \mu_3, \tag{7.1.6b}$$

$$\mathbb{E}(M_4) = \frac{n-1}{n^3} \left[3(2n-3)\mu_2^2 + (n^2 - 3n + 3)\mu_4 \right], \tag{7.1.6c}$$

$$\mathbb{E}(M_5) = \frac{(n-1)(n-2)}{n^4} \left[10(n-2)\mu_2\mu_3 + (n^2 - 2n + 2)\mu_5 \right], \tag{7.1.6d}$$

and

$$\operatorname{Var}(M_2) = \frac{(n-1)^2}{n^3} \mu_4 - \frac{(n-1)(n-3)}{n^3} \mu_2^2. \tag{7.1.7}$$

Theorem 7.1.3

Let

$$G_1 = \frac{n^2}{(n-1)(n-2)} \frac{M_3}{S^3} = \frac{\sqrt{n(n-1)}}{n-2} \frac{M_3}{M_2^{3/2}}, \tag{7.1.8a}$$

$$G_2 = \frac{n-1}{(n-2)(n-3)} \left[(n+1) \frac{M_4}{M_2^2} - 3(n-1) \right] + 3 \tag{7.1.8b}$$

then $\mathbb{E}(G_1) = \gamma_1$ is the unbiased estimator of skewness, and $\mathbb{E}(G_2) = \gamma_2$ is the unbiased estimator of kurtosis.

Theorem 7.1.4

Let $X_1, \ldots, X_n$ denotes an IID random sample, then for any constant c,

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (X_i - c)^2 - n(\bar{X} - c)^2. \tag{7.1.9}$$

Proof For any constant c, since $\bar{X} - c = \frac{1}{n}\sum_{i=1}^{n}(X_i - c)$,

$$\begin{aligned}
\sum_{i=1}^{n} (X_i - \bar{X})^2 &= \sum_{i=1}^{n} \Big((X_i - c) - (\bar{X} - c) \Big)^2 \\
&= \sum_{i=1}^{n} \Big((X_i - c)^2 - 2(X_i - c)(\bar{X} - c) + (\bar{X} - c)^2 \Big) \\
&= \sum_{i=1}^{n} (X_i - c)^2 - 2(\bar{X} - c) \sum_{i=1}^{n} (X_i - c) + n(\bar{X} - c)^2 \\
&= \sum_{i=1}^{n} (X_i - c)^2 - n(\bar{X} - c)^2.
\end{aligned}$$

□

Theorem 7.1.5

Let $X_1, \ldots, X_n$ denotes a random sample from $f(x)$ with $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

- Let
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7.1.10}$$
be a statistic called the **sample mean**, then
$$\mathbb{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \sigma^2/n. \tag{7.1.11}$$
- Let
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) \tag{7.1.12}$$
be a statistic called the **sample variance**, then
$$\mathbb{E}(S^2) = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \mu_2^2 \right). \tag{7.1.13}$$

Proof For the sample mean:

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(X_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(X\right) = \mu.$$

Since $X_1, \ldots, X_n$ are i.i.d. random variables, we have $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, and

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\left(X\right) = \frac{\sigma^2}{n}.$$

For the sample variance:

$$\begin{aligned}\mathbb{E}(S^2) &= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right]\\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}^2)\right]\\ &= \frac{1}{n-1}\left[n(\mu^2+\sigma^2) - n\left(\mu^2+\frac{\sigma^2}{n}\right)\right] = \sigma^2.\end{aligned}$$

Let $Z_i = X_i - \mu$ for $i = 1, \ldots, n$ so that $\mathbb{E}(Z_i) = 0$. Since $\text{Var}(S^2) = \mathbb{E}(S^4) - [\mathbb{E}(S^2)]^2$, we derive an expression of $\mathbb{E}(S^4)$. we can write

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}Z_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}Z_i\right)^2\right],$$

by squaring,

$$S^4 = \frac{1}{(n-1)^2}\left[\left(\sum_{i=1}^{n}Z_i^2\right)^2 - \frac{2}{n}\left(\sum_{i=1}^{n}Z_i^2\right)\left(\sum_{i=1}^{n}Z_i\right)^2 + \frac{1}{n^2}\left(\sum_{i=1}^{n}Z_i\right)^4\right].$$

Since $Z_1, \ldots, Z_n$ are independent, we have $\mathbb{E}(Z_iZ_j) = \mathbb{E}(Z_i)\mathbb{E}(Z_j) = 0$ for $i \neq j$, and

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n}Z_i^2\right)^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}Z_i^4 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}Z_i^2Z_j^2\right]\\ &= n\mu_4 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}(Z_i^2)\mathbb{E}(Z_j^2) = n\mu_4 + n(n-1)\mu_2^2,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n}Z_i^2\right)\left(\sum_{i=1}^{n}Z_i\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n}Z_i^2\right)\left(\sum_{i=1}^{n}Z_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}Z_iZ_j\right)\right]\\ &= n\mu_4 + n(n-1)\mu_2^2,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n}Z_i\right)^4\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n}Z_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}Z_iZ_j\right)^2\right]\\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n}Z_i^2\right)^2\right] + 4\mathbb{E}\left[\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}Z_iZ_j\right)^2\right]\\ &= n\mu_4 + 3n(n-1)\mu_2^2,\end{aligned}$$

thus

$$\begin{aligned}\mathbb{E}(S^4) &= \frac{1}{n^2(n-1)^2}\left[n(n^2-2n+1)\mu_4 + n(n-1)(n^2-2n+3)\mu_2^2\right]\\ &= \frac{1}{n}\mu_4 + \frac{(n-3)(n+1)}{n(n-1)}\mu_2^2,\end{aligned}$$

and

$$\operatorname{Var}(S^2) = \mathbb{E}(S^4) - [\mathbb{E}(S^2)]^2 = \mathbb{E}(S^4) - \mu_2^2 = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right). \qquad \square$$

7.2 Sampling Distributions

Theorem 7.2.1: Linear Combinations of Normal Variables

If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$; $i = 1, \ldots, n$ denote independent normal variables, then

$$Y = \sum_{i=1}^{n} a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i,\ \sum_{i=1}^{n} a_i^2\sigma_i^2\right). \tag{7.2.1}$$

Proof

$$\begin{aligned}M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(a_i t) = \prod_{i=1}^{n}\exp(a_i\mu_i t + a_i^2\sigma_i^2 t^2/2)\\ &= \exp\left(t\sum_{i=1}^{n} a_i\mu_i + t^2\sum_{i=1}^{n} a_i^2\sigma_i^2 \Big/ 2\right)\end{aligned}$$

which is the MGF of a normal variable with mean $\sum_{i=1}^{n} a_i\mu_i$ and variance $\sum_{i=1}^{n} a_i^2\sigma_i^2$. □

Theorem 7.2.2: Noncentral Chi-Squared Distribution

If $X_1, \ldots, X_n$ be n independent, normally distributed random variables with means μ_i and unit variances, $X_i \sim \mathcal{N}(\mu_i, 1)$, then the random variable

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n, \delta), \tag{7.2.2}$$

is referred to as **noncentral chi-squared distribution** with k degrees of freedom and noncentrality parameter

$$\delta = \sum_{i=1}^{n} \mu_i^2. \tag{7.2.3}$$

The probability density function of Y is

$$f(x) = \sum_{k=0}^{\infty} \frac{e^{-\delta/2}(\delta/2)^k}{k!}\,\frac{x^{(n+2k)/2-1}e^{-x/2}}{2^{(n+2k)/2}\Gamma[(n+2k)/2]}, \quad x \geq 0. \tag{7.2.4}$$

Its MGF is given by

$$M_Y(t) = (1-2t)^{-n/2} \exp\left(\frac{\delta t}{1-2t}\right). \tag{7.2.5}$$

If $\delta = 0$, then $Y \sim \chi^2(n)$, which is the **chi-squared distribution**, with PDF given by

$$f(x) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}, \quad x \geq 0. \tag{7.2.6}$$

Proof Let $Z_i \sim \mathcal{N}(0,1)$, then $X_i = Z_i + \mu_i$. The MGF of X_i^2 is given by

$$\begin{aligned} M_{X_i^2}(t) &= \mathbb{E}\left[\exp\left(tX_i^2\right)\right] = \mathbb{E}\left[\exp\left(t(Z_i+\mu_i)^2\right)\right] \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[t(z+\mu_i)^2 - z^2/2\right] \,\mathrm{d}z \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\left(\frac{1}{2}-t\right)\left(z - \frac{2\mu_i t}{1-2t}\right)^2 + \frac{\mu_i^2 t}{1-2t}\right] \,\mathrm{d}z \\ &= (1-2t)^{-1/2} \exp\left(\frac{\mu_i^2 t}{1-2t}\right), \end{aligned}$$

and the MGF of Y is given by

$$\begin{aligned} M_Y(t) &= \mathbb{E}[\exp(tY)] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(tX_i^2\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(t(Z_i+\mu_i)^2\right)\right] \\ &= \prod_{i=1}^n (1-2t)^{-1/2} \exp\left(\frac{\mu_i^2 t}{1-2t}\right) = (1-2t)^{-n/2} \exp\left(\frac{\delta t}{1-2t}\right). \end{aligned}$$ □

Theorem 7.2.3

If $Y_i \sim \chi^2(\nu_i)$; $i = 1, \ldots, n$ are independent chi-square variables, then

$$V = \sum_{i=1}^n Y_i \sim \chi^2\left(\sum_{i=1}^n \nu_i\right). \tag{7.2.7}$$

Proof

$$M_V(t) = (1-2t)^{-\nu_1/2} \cdots (1-2t)^{-\nu_n/2} = (1-2t)^{-\sum \nu_i/2}$$

which is the MGF of $\chi^2(-\sum \nu_i)$. □

Theorem 7.2.4

If $J \sim \text{Poisson}(\delta)$, then $\chi^2(\nu + 2J) \sim \chi^2(\nu, \delta)$.

Theorem 7.2.5

If $V \sim \chi^2(\nu, \delta)$, then

$$\frac{V - (\nu + \delta)}{\sqrt{2(\nu + 2\delta)}} \xrightarrow{d} \mathcal{N}(0,1)$$

as either $\nu \to \infty$ or $\delta \to \infty$.

Theorem 7.2.6: Noncentral t-Distribution

If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(\nu)$, and if Z and V are independent, then the distribution of

$$T = \frac{Z + \mu}{\sqrt{V/\nu}} \sim t(\nu, \mu) \tag{7.2.8}$$

is referred to as **noncentral t-distribution** with ν degrees of freedom and noncentrality parameter μ. The probability density function of T is

$$f(x) = \frac{\nu^{\nu/2}}{\sqrt{\pi}\Gamma(\nu/2)} \frac{\exp(-\mu^2/2)}{(\nu + x^2)^{(\nu+1)/2}} \sum_{k=0}^{\infty} \Gamma[(\nu + k + 1)/2] \frac{\mu^k}{k!} \left(\frac{2x^2}{\nu + x^2}\right)^{k/2}, \quad -\infty < x < \infty. \tag{7.2.9}$$

If $\mu = 0$, then $T \sim t(\nu)$, which is the **Student's t-distribution**, with PDF given by

$$f(x) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < x < \infty. \tag{7.2.10}$$

Theorem 7.2.7: Noncentral F-Distribution

If $V_1 \sim \chi^2(\nu_1, \delta)$ and $V_2 \sim \chi^2(\nu_2)$ are independent, then the random variable

$$X = \frac{V_1/\nu_1}{V_2/\nu_2} \sim F(\nu_1, \nu_2, \delta) \tag{7.2.11}$$

is referred to as **noncentral F-distribution** with ν_1 and nu_2 degrees of freedom and noncentrality parameter δ. The probability density function of X is

$$f(x) = \sum_{k=0}^{\infty} \frac{e^{-\delta/2}(\delta/2)^k}{k! B(\nu_1/2 + k, \nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2+k} \left(\frac{\nu_2}{\nu_2 + \nu_1 x}\right)^{(\nu_1+\nu_2)/2+k} x^{\nu_1/2-1+k}, \quad x \geq 0. \tag{7.2.12}$$

If $\delta = 0$, $X \sim F(\nu_1, \nu_2)$, which is the **Snedecor's F-distribution**, with PDF given by

$$f(x) = \frac{1}{B(\nu_1/2, \nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-(\nu_1+\nu_2)/2} x^{\nu_1/2-1}, \quad x \geq 0. \tag{7.2.13}$$

Here, B is the beta function: $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x + y)$.

Example 7.2.1

Let $X_1, \ldots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$ be independent random Samples from populations with respective distributions $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Thus,

$$(n_1-1)S_1^2/\sigma_1^2 \sim \chi^2(n_1-1), \quad (n_2-1)S_2^2/\sigma_2^2 \sim \chi^2(n_2-1),$$

so that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

Theorem 7.2.8: Noncentral Beta-Distribution

If $V_1 \sim \chi^2(\nu_1, \delta)$ and $V_2 \sim \chi^2(\nu_2)$ are independent, then the random variable

$$X = \frac{V_1}{V_1+V_2} \sim \mathsf{BETA}(\nu_1/2, \nu_2/2, \delta) \tag{7.2.14}$$

is referred to as **noncentral beta-distribution**. The probability density function of X is

$$f(x) = \sum_{k=0}^{\infty} \frac{e^{-\delta/2}(\delta/2)^k}{k!B(\nu_1/2-1+k, \nu_2/2)} x^{\nu_1/2-1+k}(1-x)^{\nu_2-1}, \quad x \in [0,1]. \tag{7.2.15}$$

If $\delta = 0$, $X \sim \mathsf{BETA}(\nu_1/2, \nu_2/2)$, which is the **beta-distribution**, with PDF given by

$$f(x) = \frac{x^{\nu_1/2-1}(1-x)^{\nu_2-1}}{B(\nu_1/2-1, \nu_2/2)}, \quad x \in [0,1]. \tag{7.2.16}$$

7.3 Properties of Normal Sample

Theorem 7.3.1

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, and c be a constant, then

1. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ or $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, and
   $X_j - \bar{X} \sim \mathcal{N}\left(0, \dfrac{n-1}{n}\sigma^2\right)$ or $\sqrt{\dfrac{n}{n-1}} \dfrac{X_j-\bar{X}}{\sigma} \sim \mathcal{N}(0,1)$.
2. $\dfrac{(X_j-c)^2}{\sigma^2} \sim \chi^2\left(1, \dfrac{(\mu-c)^2}{\sigma^2}\right)$, $\dfrac{n}{n-1}\dfrac{(X_j-\bar{X})^2}{\sigma^2} \sim \chi^2(1)$,
   $\dfrac{n(\bar{X}-\mu)^2}{\sigma^2} \sim \chi^2(1)$, and $\displaystyle\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$,
3. $\bar{X}$ is independent from $X_i - \bar{X}$; $i = 1, \ldots, n$.
4. $\bar{X}$ and S^2 are independent.

5. $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

6. $\dfrac{\bar{X}-c}{S/\sqrt{n}} \sim t\left(n-1, \dfrac{\mu-c}{\sigma/\sqrt{n}}\right)$.

7. $\dfrac{n}{(n-1)^2}\dfrac{(X_j-\bar{X})^2}{S^2} \sim \text{BETA}\left(\dfrac{1}{2}, \dfrac{n-2}{2}\right)$, and $(X_j-\bar{X})/S \sim f(n)$, where

$$f(x;n) = \frac{\sqrt{n}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}(n-1)\Gamma\left(\frac{n-2}{2}\right)}\left[1-\frac{nx^2}{(n-1)^2}\right]^{n/2-2}, \quad 0<|x|<\frac{n-1}{\sqrt{n}}, \tag{7.3.1}$$

and $\lim_{n\to\infty} f(n) = \mathcal{N}(0,1)$.

Proof 1). It follows from Theorem 7.2.1. Note $X_j - \bar{X} = \frac{n-1}{n}X_j - \sum_{i\neq j} X_i$.

2). It follows from Theorem 7.2.2 and Theorem 7.2.3.

3). The joint density of $X_1, \ldots, X_n$ is

$$f_{\boldsymbol{X}}(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu}{\sigma}\right)^2\right].$$

Consider the joint transformation and the inverse transformation:

$$y_i = \begin{cases} \bar{x}, & i=1, \\ x_i-\bar{x}, & i=2,\ldots,n, \end{cases} \qquad x_i = \begin{cases} y_1 - \sum_{i=2}^n y_i, & i=1, \\ y_1+y_i, & i=2,\ldots,n. \end{cases}$$

Thus, the joint density of $Y_1,\ldots,Y_n$ is

$$\begin{aligned} f_{\boldsymbol{Y}}(y_1,\ldots,y_n) &= \frac{1}{(2\pi)^{n/2}\sigma^n}\left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right| \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right] \\ &= \frac{n}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)\right] \\ &= \frac{n}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\left[\left(-\sum_{i=2}^n y_i\right)^2 + \sum_{i=2}^n y_i^2 + n(y_1-\mu)^2\right]\right\} \end{aligned}$$

Therefore, the joint density function factors into the marginal density function of y_1 times a function of $y_2,\ldots,y_n$ only, which shows that $Y_1 = \bar{X}$ is independent from $Y_i = X_i - \bar{X}; i = 2,\ldots,n$. Because $X_1 - \bar{X} = -\sum_{i=2}^n (X_i - \bar{X})$, it follows that $\bar{X}$ and $X_1 - \bar{X}$ also are independent.

4). It follows from 3, because S^2 is a function only of the $X_i - \bar{X}; i = 1,\ldots,n$.

5). Since

$$\underbrace{\sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^2}}_{V_1} = \underbrace{\frac{(n-1)S^2}{\sigma^2}}_{V_2} + \underbrace{\frac{n(\bar{X}-\mu)^2}{\sigma^2}}_{V_3},$$

$V_1 \sim \chi^2(n)$, $V_3 \sim \chi^2(1)$, and V_2 and V_3 are independent, so $V_2 \sim \chi^2(n-1)$.

6). For any constant c,

$$\frac{\bar{X}-c}{S/\sqrt{n}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} + \dfrac{\mu-c}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} \sim t\left(n-1, \frac{\mu-c}{\sigma/\sqrt{n}}\right).$$

7). Without loss of generality, we consider the case $j=1$. Let $\boldsymbol{Q}$ be an orthogonal matrix given by

$$\boldsymbol{Q} = \begin{bmatrix} - & \boldsymbol{q}_1^\top & - \\ - & \boldsymbol{q}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{q}_n^\top & - \end{bmatrix}$$

where

$$\boldsymbol{q}_1^\top = \sqrt{\frac{n}{n-1}}\left(1-\frac{1}{n}, -\frac{1}{n}, \dots, -\frac{1}{n}\right), \quad \boldsymbol{q}_2^\top = \frac{1}{n}(1,1,\dots,1).$$

Let $u_i = (x_i-\mu)/\sigma$ for $i=1,\dots,n$ and $\boldsymbol{z} = \boldsymbol{Q}\boldsymbol{u}$, then

$$\begin{aligned} \operatorname{Cov}(Z_i, Z_k) &= \operatorname{Cov}\left(\sum_{j=1}^n q_{ij}U_j, \sum_{l=1}^n q_{kl}U_l\right) = \sum_{j=1}^n\sum_{l=1}^n q_{ij}q_{kl}\operatorname{Cov}(U_j,U_l) \\ &= \sum_{j=1}^n\sum_{l=1}^n q_{ij}q_{kl}\delta_{jl} = \sum_{j=1}^n q_{ij}q_{kj} = \delta_{ik}, \end{aligned}$$

so Z_i and Z_k are independent for $i \neq k$. And

$$Z_1 = \sqrt{\frac{n}{n-1}}\frac{X_1-\bar{X}}{\sigma}, \quad Z_2 = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}, \quad \frac{(n-1)S^2}{\sigma^2} = \sum_{i\neq 2} Z_i^2.$$

Since $Z_1^2 \sim \chi^2(1)$ and $\sum_{i\neq 2} Z_i^2 \sim \chi^2(n-1)$, we have $\sum_{i=3}^n Z_i^2 \sim \chi^2(n-2)$ and

$$V = \frac{n}{(n-1)^2}\frac{(X_1-\bar{X})^2}{S^2} = \frac{Z_1^2}{Z_1^2+\sum_{i=3}^n Z_i^2} \sim \text{BETA}\left(\frac{1}{2}, \frac{n-2}{2}\right).$$

The PDF of V is given by

$$f\left(v; \frac{1}{2}, \frac{n-2}{2}\right) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} v^{-1/2}(1-v)^{n/2-2}, \quad 0 < v < 1.$$

Now making the transformation $W = \frac{n-1}{\sqrt{n}}\sqrt{V} = |(X_1-\bar{X})/S|$, then $V = \frac{n}{(n-1)^2}W^2$, and

$$\begin{aligned} f_W(w) &= \frac{2nw}{(n-1)^2} \cdot \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}\left[\frac{n}{(n-1)^2}w^2\right]^{-1/2}\left[1-\frac{n}{(n-1)^2}w^2\right]^{n/2-2} \\ &= \frac{2\sqrt{n}}{n-1}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}\left[1-\frac{nw^2}{(n-1)^2}\right]^{n/2-2}, \quad 0 < w < \frac{n-1}{\sqrt{n}}. \end{aligned}$$

Eventually, we get the PDF of $(X_1 - \bar{X})/S$ given by

$$f(x;\, n) = \frac{\sqrt{n}}{n-1} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)} \left[1 - \frac{nx^2}{(n-1)^2}\right]^{n/2-2}, \quad 0 < |x| < \frac{n-1}{\sqrt{n}}.$$

Chapter 8

Point Estimation

8.1 Introduction

Definition 8.1.1: Estimator & Estimate

A statistic, $T = \mathscr{T}(X_1, X_2, \ldots, X_n)$, that is used to estimate the value of $\tau(\theta)$ is called an **estimator** of $\tau(\theta)$, and an observed value of the statistic, $t = \mathscr{T}(x_1, x_2, \ldots, x_n)$, is called an **estimate** of $\tau(\theta)$.

8.2 Some Methods of Estimation

Method of Moments

Definition 8.2.1: Sample Moments

If $X_1, \ldots, X_n$ is a random sample from $f(x; \theta_1, \ldots, \theta_k)$, the first k **sample moments** are given by

$$M'_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j, \quad j = 1, 2, \ldots, k. \tag{8.2.1}$$

The method of moments principle is to choose as estimators of the parameters $\theta_1, \ldots, \theta_k$ the values $\hat{\theta}_1, \ldots, \hat{\theta}_k$ that render the population moments equal to the sample moments. In other words, $\hat{\theta}_1, \ldots, \hat{\theta}_k$ are solutions of the equations

$$M'_j = \mu'_j(\hat{\theta}_1, \ldots, \hat{\theta}_k), \quad j = 1, 2, \ldots, k. \tag{8.2.2}$$

Method of Maximum Likelihood

Definition 8.2.2: Likelihood Function

The joint density function of n random variables $X_1, \ldots, X_n$ evaluated at $x_1, \ldots, x_n$, say $f(x_1, \ldots, x_n; \theta)$, is referred to as a **likelihood function**. For fixed $x_1, \ldots, x_n$, the likelihood function is a function of θ and often is denoted by $L(\theta)$.

If $X_1, \cdots, X_n$ represents a random sample of size n from $f(x;\theta)$, then

$$L(\theta) = f(x_1;\theta) \cdots f(x_n;\theta). \tag{8.2.3}$$

Definition 8.2.3: MLE

Let $L(\theta) = f(x_1, \ldots, x_n;\, \theta)$, $\theta \in \Omega$, be the joint PDF of $X_1, \ldots, X_n$. For a given set of observations, $(x_1, \ldots, x_n)$, a value $\hat{\theta}$ in Ω at which $L(\theta)$ is a maximum is called a **maximum likelihood estimate** (MLE) of θ. That is, $\hat{\theta}$ is a value of θ that satisfies

$$f(x_1, \ldots, x_n;\, \hat{\theta}) = \max_{\theta \in \Omega} f(x_1, \ldots, x_n;\, \theta). \tag{8.2.4}$$

Notice that if each set of observations $(x_1, \ldots, x_n)$ corresponds to a unique value $\hat{\theta}$, then this procedure defines a function, $\hat{\theta} = \mathscr{T}(x_1, \ldots, x_n)$. This same function, when applied to the random sample, $\hat{\theta} = \mathscr{T}(X_1, \ldots, X_n)$, is called the **maximum likelihood estimator**, also denoted MLE. Usually, the same notation, $\hat{\theta}$, is used for both the ML estimate and the ML estimator.

In most applications, $L(\theta)$ represents the joint PDF of a random sample, although the maximum likelihood principle also applies to other cases such as sets of order statistics.

If Ω is an open interval , and if $L(\theta)$ is differentiable and assumes a maximum on Ω, then the MLE will be a solution of the equation (ma ximum likelihood equation)

$$\frac{\mathrm{d}}{\mathrm{d}\theta} L(\theta) = 0. \tag{8.2.5}$$

Strictly speaking, if one or more solutions of equation (8.2.5) exist, then it should be verified which, if any, maximize $L(\theta)$. Note also that any value of 0 that maximizes $L(\theta)$ also will maximize the log-likelihood, $\ln L(\theta)$, so for computational convenience the alternate form of the maximum likelihood equation,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln L(\theta) = 0. \tag{8.2.6}$$

often will be used.

Theorem 8.2.1: Invariance Property

If $\hat{\theta}$ is the MLE of θ and if $u(\theta)$ is a function of θ, then $u(\hat{\theta})$ is an MLE of $u(\theta)$.

The definitions of likelihood function and maximum likelihood estimator can be applied in the case of more than one unknow n parameter if $\boldsymbol{\theta}$ represents a vector of parameters, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Although n could, in general, be almost any sort of k-dimensional set, in most examples it is a Cartesian product of k intervals. When n is of this form and if the partial derivatives of $L(\theta_1, \ldots, \theta_k)$ exist, and the MLEs do not occur on the boundary of Ω, then the MLEs will be solutions of the simultaneous equations

$$\frac{\partial}{\partial \theta_j} \ln L(\theta_1, \ldots, \theta_k) = 0, \quad j = 1, \ldots, k. \tag{8.2.7}$$

These are called the **maximum likelihood equations**, and the solutions are denoted by $\hat{\theta}_1, \ldots, \hat{\theta}_k$. As in the one-parameter case, it generally is necessary to verify that the solutions of the ML equations maximize $L(\theta_1, \ldots, \theta_k)$.

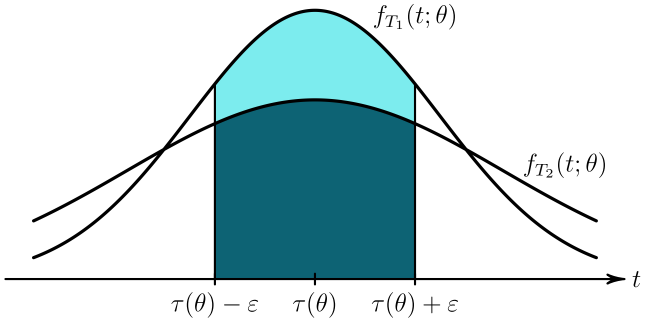

Figure 8.1: The concept of "more concentrated"

8.3 Criteria for Evaluating Estimators

Definition 8.3.1: Unbiased & Biased Estimator

An estimator T is said to be an **unbiased estimator** of $\tau(\theta)$ if $\mathbb{E}(T) = \tau(\theta)$ for all $\theta \in \Omega$. Otherwise, we say that T is a **biased estimator** of $\tau(\theta)$.

If an unbiased estimator is used to assign a value of $\tau(\theta)$, then the correct value of $\tau(\theta)$ may not be achieved by any given estimate, T, but the "average" value of T will be $\tau(\theta)$.

It might be reasonable to say that T_1 is **more concentrated** than T_2 about $\tau(\theta)$ if

$$\mathbb{P}[\tau(\theta) - \varepsilon < T_1 < \tau(\theta) + \varepsilon] \geq \mathbb{P}[\tau(\theta) - \varepsilon < T_2 < \tau(\theta) + \varepsilon] \tag{8.3.1}$$

for all $\varepsilon > 0$, and that an estimator is **most concentrated** if it is more concentrated than any other estimator.

The idea of a more concentrated estimator is illustrated in figure 8.1, which shows the PDFs of two estimators T_1 and T_2.

It is not clear how to obtain an estimator that is most concentrated, but some other concepts will be discussed that may partially achieve this goal. For example, if T is an unbiased estimator of $\tau(\theta)$, it follows from the Chebyshev Inequality that

$$\mathbb{P}[\tau(\theta) - \varepsilon < T < \tau(\theta) + \varepsilon] \geq 1 - \mathrm{Var}(T)/\varepsilon^2 \tag{8.3.2}$$

for all $\varepsilon > 0$. This suggests that for unbiased estimators, one with a smaller variance will tend to be more concentrated and thus may be preferable.

Uniformly Minimum Variance Unbiased Estimators

Definition 8.3.2: UMVUE

Let $X_1, \ldots, X_n$ be a random sample of size n from $f(x;\theta)$. An estimator T^* of $\tau(\theta)$ is called **uniformly minimum variance unbiased estimators** (UMVUE) of $\tau(\theta)$ if

1. T^* is unbiased for $\tau(\theta)$, and

2. for any other unbiased estimator T of $\tau(\theta)$, $\mathrm{Var}(T^*) \le \mathrm{Var}(T)$ for all $\theta \in \Omega$.

Definition 8.3.3

Let $T = \mathscr{T}(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$, then under smoothness assumptions on $f(x_1, \ldots, x_n; \theta)$, the **score function** is defined to be

$$\mathscr{S}(x_1, \cdots, x_n; \theta) = \frac{\partial}{\partial\theta} \ln f(x_1, \cdots, x_n; \theta). \tag{8.3.3}$$

The **Fisher information** is defined by the two equivalent expressions

$$I_n(\theta) := \mathrm{Var}_\theta[\mathscr{S}(X_1, \cdots, X_n; \theta)], \tag{8.3.4}$$

where Var_θ denote variance with respect to $(X_1, \ldots, X_n) \sim f(x_1, \cdots, x_n; \theta)$.

Theorem 8.3.1

Let $S = \mathscr{S}(X_1, \cdots, X_n; \theta)$ be a random variable. Under regularity conditions,

$$\mathbb{E}_\theta(S) = 0, \quad I_n(\theta) = \mathrm{Var}_\theta(S) = \mathbb{E}_\theta(S^2) = -\mathbb{E}_\theta\left(\frac{\partial S}{\partial\theta}\right). \tag{8.3.5}$$

Proof Since

$$\mathscr{S}(x_1, \cdots, x_n; \theta) = \frac{\partial}{\partial\theta} \ln f(x_1, \cdots, x_n; \theta) = \frac{1}{f(x_1, \cdots, x_n; \theta)} \frac{\partial}{\partial\theta} f(x_1, \cdots, x_n; \theta),$$

we have

$$\begin{aligned}\mathbb{E}_\theta(S) &= \int \cdots \int \mathscr{S}(x_1, \cdots, x_n; \theta) f(x_1, \cdots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int \cdots \int \frac{\partial}{\partial\theta} f(x_1, \cdots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \frac{\mathrm{d}}{\mathrm{d}\theta} \int \cdots \int f(x_1, \cdots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = \frac{\mathrm{d}}{\mathrm{d}\theta} 1 = 0.\end{aligned}$$

Thus, $I_n(\theta) := \mathrm{Var}_\theta(S) = \mathbb{E}_\theta(S^2) - [\mathbb{E}_\theta(S)]^2 = \mathbb{E}_\theta(S^2)$.

Let f denotes $f(x_1, \cdots, x_n; \theta)$, then

$$\frac{\partial^2}{\partial\theta^2} \ln f = \frac{\partial}{\partial\theta}\left(\frac{1}{f}\frac{\partial f}{\partial\theta}\right) = -\frac{1}{f^2}\left(\frac{\partial f}{\partial\theta}\right)^2 + \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2} = -\left(\frac{\partial}{\partial\theta}\ln f\right)^2 + \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2},$$

and

$$\begin{aligned}\mathbb{E}\left(\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\right) &= \int \cdots \int \frac{\partial^2}{\partial\theta^2} f(x_1, \ldots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \int \cdots \int f(x_1, \ldots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} 1 = 0,\end{aligned}$$

so

$$\mathbb{E}_\theta(S^2) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f\right] = -\mathbb{E}_\theta\left(\frac{\partial S}{\partial\theta}\right).$$

□

Theorem 8.3.2

Let $X_1, \ldots, X_n$ be a random sample of size n from $f(x;\theta)$, so that

$$I_n(\theta) = nI(\theta), \tag{8.3.6}$$

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right]. \tag{8.3.7}$$

where $\mathbb{E}_\theta$ denote expectation with respect to $X \sim f(x;\theta)$.

Theorem 8.3.3: Cramér-Rao Lower Bound

Let $T = \mathscr{T}(X_1, \ldots, X_n)$ be an unbiased estimator of $\tau(\theta)$, then under smoothness assumptions on $f(x_1, \ldots, x_n; \theta)$,

$$\operatorname{Var}(T) \geq \frac{[\tau'(\theta)]^2}{I_n(\theta)}. \tag{8.3.8}$$

Proof According to the definition 8.3.3, we define $S = \mathscr{S}(X_1, \ldots, X_n; \theta)$, then $\mathbb{E}(S) = 0$. Since $T = \mathscr{T}(X_1, \ldots, X_n)$ is unbiased for $\tau(\theta)$,

$$\tau(\theta) = \mathbb{E}(T) = \int \cdots \int \mathscr{T}(x_1, \ldots, x_n) f(x_1, \ldots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

If we differentiate with respect to θ, then

$$\begin{aligned}\tau'(\theta) &= \int \cdots \int \mathscr{T}(x_1, \ldots, x_n) \frac{\partial}{\partial\theta} f(x_1, \ldots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int \cdots \int \mathscr{T}(x_1, \ldots, x_n; \theta) \mathscr{S}(x_1, \ldots, x_n; \theta) f(x_1, \ldots, x_n; \theta)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \mathbb{E}(TS).\end{aligned}$$

Since $\mathbb{E}(S) = 0$, we have $\operatorname{Var}(S) = \mathbb{E}(S^2)$ and $\operatorname{Cov}(T, S) = \mathbb{E}(TS)$. Because $(\operatorname{Cov}(T,S))^2 \leq \operatorname{Var}(T)\operatorname{Var}(S)$, and consequently

$$\operatorname{Var}(T) \geq \frac{[\operatorname{Cov}(T,S)]^2}{\operatorname{Var}(S)} = \frac{[\mathbb{E}(TS)]^2}{\mathbb{E}(S^2)} = \frac{[\tau'(\theta)]^2}{I_n(\theta)}. \qquad \square$$

Definition 8.3.4: Efficiency

The relative efficiency of an unbiased estimator T of $\tau(\theta)$ to another unbiased estimator T^* of $\tau(\theta)$ is given by

$$\mathsf{re}(T, T^*) = \frac{\operatorname{Var}(T^*)}{\operatorname{Var}(T)}. \tag{8.3.9}$$

An unbiased estimator T^* of $\tau(\theta)$ is said to be **efficient** if $\mathsf{re}(T, T^*) \leq 1$ for all unbiased estimators T of $\tau(\theta)$, and all $\theta \in \Omega$. The efficiency of an unbiased estimator T of $tau(\theta)$ is given by

$$\mathsf{e}(T) = \mathsf{re}(T, T^*), \tag{8.3.10}$$

if T^* is an efficient estimator of $\tau(\theta)$.

Definition 8.3.5: Bias & Mean Squared Error

If T is an estimator of $\tau(\theta)$, then the **bias** is given by

$$\mathsf{bias}(T) = \mathbb{E}(T) - \tau(\theta), \tag{8.3.11}$$

and the **mean squared error** (MSE) of T is given by

$$\mathsf{MSE}(T) = \mathbb{E}[(T - \tau(\theta))^2]. \tag{8.3.12}$$

Theorem 8.3.4

If T is an estimator of $\tau(\theta)$, then

$$\mathsf{MSE}(T) = \mathrm{Var}(T) + (\mathsf{bias}(T))^2. \tag{8.3.13}$$

Proof

$$\begin{aligned}
\mathsf{MSE}(T) &= \mathbb{E}[(T - \tau(\theta))^2] \\
&= \mathbb{E}[(T - \mathbb{E}(T) + \mathbb{E}(T) - \tau(\theta))^2] \\
&= \mathbb{E}[(T - \mathbb{E}(T))^2] + 2\mathbb{E}[T - \mathbb{E}(T)] \cdot (\mathbb{E}(T) - \tau(\theta)) + (\mathbb{E}(T) - \tau(\theta))^2 \\
&= \mathrm{Var}(T) + (\mathsf{bias}(T))^2.
\end{aligned}$$

□

8.4 Large-Sample Properties

Definition 8.4.1: Simple Consistency

Let $\{T_n\}$ be a sequence of estimators of $\tau(\theta)$. Theses estimators are said to be **consistent** estimators of $\tau(\theta)$ if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|T_n - \tau(\theta)| < \varepsilon] = 1, \quad \forall\, \theta \in \Omega. \tag{8.4.1}$$

Definition 8.4.2: MSE Consistency

If $\{T_n\}$ is a sequence of estimators of $\tau(\theta)$, then they are called **mean squared error consistent** if

$$\lim_{n\to\infty} \mathsf{MSE}(T_n) = \lim_{n\to\infty} \mathbb{E}[(T_n - \tau(\theta))^2] = 0, \quad \forall\, \theta \in \Omega. \tag{8.4.2}$$

Definition 8.4.3: Asymptotic Unbiased

A sequence $\{T_n\}$ is said to be **asymptotically unbiased** for $\tau(\theta)$ if

$$\lim_{n\to\infty} \mathbb{E}(T_n) = \tau(\theta), \quad \forall\, \theta \in \Omega. \tag{8.4.3}$$

Theorem 8.4.1

A sequence $\{T_n\}$ of estimators of $\tau(\theta)$ is mean squared error consistent if and only if it is asymptotically unbiased and $\lim_{n\to\infty} \text{Var}(T_n) = 0$.

Proof Because

$$\mathsf{MSE}(T_n) = \text{Var}(T_n) + (\mathbb{E}(T_n) - \tau(\theta))^2,$$

and both terms on the right are nonnegative, $\mathsf{MSE}(T_n) \to 0$ implies both $\text{Var}(T_n) \to 0$ and $\mathbb{E}(T_n) \to \tau(\theta)$. The converse is obvious. □

Theorem 8.4.2

If a sequence $\{T_n\}$ is mean squared error consistent, it also is simply consistent.

Proof This follows from the Markov inequality (2.4.8), with $X = T_n - \tau(\theta)$, $r = 2$ and $c = \varepsilon$, so that

$$\mathbb{P}[|T_n - \tau(\theta)| < \varepsilon] \geq 1 - \mathsf{MSE}(T_n)/\varepsilon^2$$

which approaches 1 as $n \to \infty$. □

Theorem 8.4.3

If $\{T_n\}$ is simply consistent for $\tau(\theta)$ and if $g(t)$ is continuous at each value of $\tau(\theta)$, then $g(T_n)$ is simply consistent for $g(\tau(\theta))$.

Proof This follows immediately from Theorem 6.1.2 with $Y_n = T_n$ and $c = \tau(\theta)$.

Definition 8.4.4: Asymptotic Efficiency

Let $\{T_n\}$ and $\{T_n^*\}$ be two asymptotically unbiased sequences of estimators for $\tau(\theta)$. The **asymptotic relative efficiency** of T_n relative to T_n^* is given by

$$\mathsf{are}(T_n, T_n^*) = \lim_{n\to\infty} \frac{\text{Var}(T_n^*)}{\text{Var}(T_n)}. \tag{8.4.4}$$

The sequence $\{T_n^*\}$ is said to be asymptotically efficient if $\mathsf{are}(T, T^*) \leq 1$ for all other asymptotically unbiased sequences $\{T_n\}$, and all $\theta \in \Omega$. The asymptotic efficiency of an asymptotically unbiased sequence $\{T_n\}$ is given by

$$\mathsf{ae}(T_n) = \mathsf{are}(T_n, T_n^*), \tag{8.4.5}$$

if $\{T_n^*\}$ is asymptotically efficient.

An estimator with variance of order $1/n^2$ usually is referred to as a **super efficient** estimator.

Consistency and Asymptotic Normality of the MLE

Theorem 8.4.4: Consistency and Asymptotic Normality of the MLE

Let $X_1, \dots, X_n$ be a random sample of size n from $f(x;\,\theta_0)$, and $\hat{\theta}$ be the MLE for θ_0. Suppose certain regularity conditions hold, then $\hat{\theta}$ is consistent and asymptotically normal, with

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} Z \sim \mathcal{N}(0,1), \quad \text{as } n \to \infty. \tag{8.4.6}$$

Proof **Part I (Consistency)**: Consider the log-likelihood function

$$l(\theta) = \sum_{i=1}^{n} \ln f(X_i;\,\theta),$$

where $X_1, \dots, X_n \overset{IID}{\sim} f(x;\,\theta_0)$. By the Weak Law of Large Numbers (Theorem 6.2.1), we have

$$\frac{1}{n} l(\theta) \xrightarrow{p} \mathbb{E}_{\theta_0}[\ln f(X;\,\theta)] = \int \ln f(x;\,\theta) f(x;\,\theta_0)\,\mathrm{d}x, \quad \text{as } n \to \infty.$$

Under suitable regularity conditions, this implies

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, l(\theta) \xrightarrow{p} \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{\theta_0}[\ln f(X;\,\theta)] = \theta_0.$$

Indeed, for any $\theta \in \Omega$,

$$\mathbb{E}_{\theta_0}[\ln f(X;\,\theta)] - \mathbb{E}_{\theta_0}[\ln f(X;\,\theta_0)] = \mathbb{E}_{\theta_0}\left[\ln \frac{f(X;\,\theta)}{f(X;\,\theta_0)}\right].$$

Noting that $x \mapsto \ln x$ is concave, Jensen's inequality implies $\mathbb{E}(\ln X) \le \ln \mathbb{E}(X)$ for any positive random variable X, so

$$\mathbb{E}_{\theta_0}\left[\ln \frac{f(X;\,\theta)}{f(X;\,\theta_0)}\right] \le \ln \mathbb{E}_{\theta_0}\left[\frac{f(X;\,\theta)}{f(X;\,\theta_0)}\right] = \ln \int \frac{f(x;\,\theta)}{f(x;\,\theta_0)} f(x;\,\theta_0)\,\mathrm{d}x = \ln 1 = 0.$$

So $\theta \mapsto \mathbb{E}_{\theta_0}[\ln f(X;\,\theta)]$ is maximized at $\theta = \theta_0$, which establishes consistency of $\hat{\theta}$.

Part II (Asymptotic Normality): Since $\hat{\theta}$ maximizes the log-likelihood function $l(\theta)$, we must have $l'(\hat{\theta}) = 0$. According to Mean Value Theorem, there exists θ^* between θ_0 and $\hat{\theta}$

$$0 = l'(\hat{\theta}) = l'(\theta_0) + (\hat{\theta} - \theta_0) l''(\theta^*) \quad \Rightarrow \quad \hat{\theta} - \theta_0 = -\frac{l'(\theta_0)}{l''(\theta^*)}.$$

Since $\left[\frac{\partial}{\partial\theta} \ln f(X;\,\theta)\right]_{\theta=\theta_0}$ has mean 0 and variance $I(\theta_0)$, according to Central Limit Theorem

$$\frac{l'(\theta_0)}{\sqrt{nI(\theta_0)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta} \ln f(X_i;\,\theta)\right]_{\theta=\theta_0}}{\sqrt{I(\theta_0)}/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

For the denominator, by the Law of Large Numbers and $\theta^* \xrightarrow{p} \theta_0$,

$$\frac{1}{n} l''(\theta^*) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial^2}{\partial \theta^2} \ln f(X_i;\, \theta) \right]_{\theta=\theta^*} \xrightarrow{p} \mathbb{E}_{\theta_0} \left[\frac{\partial^2}{\partial \theta^2} \ln f(X;\, \theta) \right]_{\theta=\theta_0} = -I(\theta_0).$$

Therefore, Slutsky's Theorem 6.1.5 gives $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} Z \sim \mathcal{N}(0,1)$, as $n \to \infty$. □

8.5 Bayes and Minimax Estimators

Definition 8.5.1: Loss Function

If T is an estimator of $\tau(\theta)$, then a **loss function** is any real-valued function, $L(t;\, \theta)$, such that $L(t;\, \theta) \geq 0$ for every t and $L(t;\, \theta) = 0$ when $t = \tau(\theta)$.

Definition 8.5.2: Risk Function

The **risk function** is defined to be the expected loss,

$$R_T(\theta) = \mathbb{E}[L(T;\, \theta)]. \tag{8.5.1}$$

Definition 8.5.3: Admissible Estimator

An estimator T_1 is a better estimator than T_2 if and only if $R_{T_1}(\theta) \leq R_{T_2}(\theta)$ for all $\theta \in \Omega$ and $R_{T_1}(\theta) < R_{T_2}(\theta)$ for at least one θ. An estimator T is **admissible** if and only if there is no better estimator.

Definition 8.5.4: Minimax Estimator

An estimator T_1 is a **minimax estimator** if

$$\max_{\theta} R_{T_1}(\theta) \leq \max_{\theta} R_T(\theta) \tag{8.5.2}$$

for every estimator T.

Definition 8.5.5: Bayes Risk

For a random sample from $f(x;\, \theta)$, the **Bayes risk** of an estimator T relative to a risk function $R_T(\theta)$ and PDF $p(\theta)$ is the average risk with respect to $p(\theta)$,

$$\mathbb{E}_{\theta}[R_T(\theta)] = \int_{\Omega} R_T(\theta) p(\theta)\, \mathrm{d}\theta. \tag{8.5.3}$$

Definition 8.5.6: Bayes Estimator

For a random sample from $f(x;\theta)$, the **Bayes estimator** T^* relative to the risk function $R_T(\theta)$ and PDF $p(\theta)$ is the estimator with minimum expected risk,

$$\mathbb{E}_\theta[R_{T^*}(\theta)] \le \mathbb{E}_\theta[R_T(\theta)] \tag{8.5.4}$$

for every estimator T.

Definition 8.5.7: Posterior Distribution

The conditional density of θ given the sample observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ is called the **posterior density** or **posterior PDF**, and is given by

$$f_{\theta|\boldsymbol{x}}(\theta) = \frac{f(x_1,\ldots,x_n|\theta)p(\theta)}{\displaystyle\int f(x_1,\ldots,x_n|\theta)p(\theta)\,\mathrm{d}\theta}. \tag{8.5.5}$$

The Bayes estimator is the estimator that minimizes the average risk over θ, $\mathbb{E}_\theta[R_T(\theta)]$. However,

$$\mathbb{E}_\theta[R_T(\theta)] = \mathbb{E}_\theta\{\mathbb{E}_{\boldsymbol{X}|\theta}[L(T;\theta)]\} = \mathbb{E}_{\boldsymbol{X}}\{\mathbb{E}_{\theta|\boldsymbol{X}}[L(T;\theta)]\}$$

and an estimator T that minimizes $E_{\theta|\boldsymbol{X}}[L(T;\theta)]$ for each $\boldsymbol{x}$ also minimizes the average over $\boldsymbol{X}$. Thus the Bayes estimator may be obtained by minimizing the expected loss relative to the posterior distribution.

Theorem 8.5.1

If $X_1, \ldots, X_n$ denotes a random sample from $f(x|\theta)$, then the Bayes estimator is the estimator that minimizes the expected loss relative to the posterior distribution of $\theta|\boldsymbol{x}$, $\mathbb{E}_{\theta|\boldsymbol{x}}[L(T;\theta)]$.

Theorem 8.5.2

The Bayes estimator, T, of $\tau(\theta)$ under the squared error loss function,

$$L(T;\theta) = [T - \tau(\theta)]^2, \tag{8.5.6}$$

is the conditional mean of $\tau(\theta)$ relative to the posterior distribution,

$$T = \mathbb{E}_{\theta|\boldsymbol{X}}[\tau(\theta)] = \int \tau(\theta) f_{\theta|\boldsymbol{X}}(\theta)\,\mathrm{d}\theta. \tag{8.5.7}$$

Proof

$$\mathbb{E}_{\theta|\boldsymbol{X}}[L(T;\theta)] = \mathbb{E}_{\theta|\boldsymbol{X}}[(T-\tau(\theta))^2] = \int (T-\tau(\theta))^2 f_{\theta|\boldsymbol{X}}(\theta)\,\mathrm{d}\theta,$$

is minimized when

$$0 = \frac{\mathrm{d}}{\mathrm{d}T}\mathbb{E}_{\theta|\boldsymbol{X}}[L(T;\theta)] = \int 2(T-\tau(\theta)) f_{\theta|\boldsymbol{X}}(\theta)\,\mathrm{d}\theta = 2\left(T - \int \tau(\theta) f_{\theta|\boldsymbol{X}}(\theta)\,\mathrm{d}\theta\right). \qquad \square$$

Theorem 8.5.3

The Bayes estimator, $\hat{\theta}$, of θ under loss $(a, b > 0)$

$$L(\hat{\theta};\, \theta) = \begin{cases} a|\hat{\theta} - \theta|, & \hat{\theta} - \theta \geq 0 \\ b|\hat{\theta} - \theta|, & \hat{\theta} - \theta < 0 \end{cases} \tag{8.5.8}$$

yields a quantile from the posterior distribution $f_{\theta|\boldsymbol{X}}(\theta)$.

Proof

$$\begin{aligned} \mathbb{E}_{\theta|\boldsymbol{X}}[L(\hat{\theta};\, \theta)] &= \int L(\hat{\theta};\, \theta) f_{\theta|\boldsymbol{X}}(\theta)\, \mathrm{d}\theta \\ &= \int_{-\infty}^{\hat{\theta}} a(\hat{\theta} - \theta) f_{\theta|\boldsymbol{X}}(\theta)\, \mathrm{d}\theta - \int_{\hat{\theta}}^{\infty} b(\hat{\theta} - \theta) f_{\theta|\boldsymbol{X}}(\theta)\, \mathrm{d}\theta \end{aligned}$$

is minimized when

$$0 = \frac{\mathrm{d}}{\mathrm{d}\hat{\theta}} \mathbb{E}_{\theta|\boldsymbol{X}}[L(\hat{\theta};\, \theta)] = aF_{\theta|\boldsymbol{X}}(\hat{\theta}) - b(1 - F_{\theta|\boldsymbol{X}}(\hat{\theta})) \Rightarrow F_{\theta|\boldsymbol{X}}(\hat{\theta}) = b/(a+b). \qquad \square$$

Theorem 8.5.4

The Bayes estimator, $\hat{\theta}$, of θ under absolute error loss,

$$L(\hat{\theta};\, \theta) = |\hat{\theta} - \theta| \tag{8.5.9}$$

is the median of the posterior distribution $f_{\theta|\boldsymbol{X}}(\theta)$.

Example 8.5.1

Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and assume that the variance σ^2 is known and the prior density of the mean μ is $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, i.e.

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right).$$

Since the likelihood function

$$f(\boldsymbol{x}|\mu) = \prod_{i=1}^{n} f(x_i|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right),$$

the posterior distribution

$$\begin{aligned} f(\mu|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\mu)p(\mu) \\ &\propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)\right] \\ &\propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) \\ &\propto \exp\left(-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}\right), \end{aligned}$$

i.e. $\mu|\boldsymbol{x} \sim \mathcal{N}(\mu_n, \sigma_n)$, where

$$\mu_n = \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\bar{x}, \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}.$$

Using squared error loss, the Bayes estimator of μ is given by

$$T = \mathbb{E}_{\mu|\mathbf{X}}(\mu) = \mu_n.$$

Theorem 8.5.5

If T^* is a Bayes estimator with constant risk, $R_{T^*}(\theta) = c$, then T^* is a minimax estimator.

Proof

$$\max_\theta R_{T^*}(\theta) = c = R_{T^*}(\theta) = \mathbb{E}_\theta[R_{T^*}(\theta)] \le \mathbb{E}_\theta[R_T(\theta)] \le \max_\theta R_T(\theta)$$

for every T. The first inequality is because T^* is the Bayes estimator. The second inequality is because the average of a variable is not larger than the maximum value of a variable. □

Chapter 9

Sufficiency and Completeness

9.1 Sufficiency

Definition 9.1.1: Jointly Sufficient Statistics

Let $\boldsymbol{X} = (X_1, \dots, X_n)$ have joint PDF $f(\boldsymbol{x}, \boldsymbol{\theta})$, and let $\boldsymbol{S} = (S_1, \dots, S_k)$ be a k-dimensional statistic. Then $S_1, \dots, S_k$ is a set of **jointly sufficient statistics** for $\boldsymbol{\theta}$ if for any other vector of statistics, $\boldsymbol{T}$, the conditional PDF of $\boldsymbol{T}$ give $\boldsymbol{S} = \boldsymbol{s}$, denoted by $f_{\boldsymbol{T}|\boldsymbol{S}}(\boldsymbol{t})$, does not depend on $\boldsymbol{\theta}$. In the one-dimensional case, we simply say that S is a **sufficient statistic** for θ.

Definition 9.1.2: Minimal Sufficient

A set of statistics is called a **minimal sufficient** set if the members of the set are jointly sufficient for the parameters and if they are a function of every other set of jointly sufficient statistics.

Theorem 9.1.1: Fisher-Neyman Factorization Criterion

If $X_1, \dots, X_n$ have joint PDF $f(x_1, \dots, x_n; \boldsymbol{\theta})$, and if $\boldsymbol{S} = (S_1, \dots, S_k)$, then $S_1, \dots, S_k$ are jointly sufficient for $\boldsymbol{\theta}$ if and only if

$$f(x_1, \dots, x_n; \boldsymbol{\theta}) = f_{\boldsymbol{s}}(\boldsymbol{s}; \boldsymbol{\theta}) f_{\boldsymbol{X}|\boldsymbol{s}}(x_1, \dots, x_n), \tag{9.1.1}$$

where $f_{\boldsymbol{s}}(\boldsymbol{s}; \boldsymbol{\theta})$ is the PDF of $\boldsymbol{S}$ and $f_{\boldsymbol{X}|\boldsymbol{s}}(x_1, \dots, x_n)$ is the conditional PDF of $\boldsymbol{X} = (X_1, \dots, X_n)$ given $\boldsymbol{S} = \boldsymbol{s}$.

Theorem 9.1.2

If $S_1, \dots, S_k$ are jointly sufficient for $\boldsymbol{\theta}$ and if $\hat{\boldsymbol{\theta}}$ is a unique MLE of $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}$ is a function of $\boldsymbol{S} = (S_1, \dots, S_k)$.

Proof By the factorization criterion,

$$L(\boldsymbol{\theta}) = f(x_1, \dots, x_n; \boldsymbol{\theta}) = f_{\boldsymbol{s}}(\boldsymbol{s}; \boldsymbol{\theta}) f_{\boldsymbol{X}|\boldsymbol{s}}(x_1, \dots, x_n)$$

which means that value that maximizes the likelihood function must depend on $\boldsymbol{s}$, say $\hat{\boldsymbol{\theta}} = \mathscr{T}(\boldsymbol{s})$. If the MLE is unique, this defines a function of $\boldsymbol{s}$. □

Theorem 9.1.3

If is sufficient for θ, then any Bayes estimator will be a function of S.

Proof According to the factorization criterion, the posterior density is

$$f_{\theta|\boldsymbol{x}}(\theta) = \frac{f(x_1, \ldots, x_n; \theta)p(\theta)}{\int f(x_1, \ldots, x_n; \theta)p(\theta)\,\mathrm{d}\theta} = \frac{f_s(s; \theta)p(\theta)}{\int f_s(s; \theta)p(\theta)\,\mathrm{d}\theta}.$$ □

Theorem 9.1.4

If $X_1, \ldots, X_n$ is a random sample from a continuous distribution with PDF $f(x; \boldsymbol{\theta})$, then the order statistic form a jointly sufficient set for $\boldsymbol{\theta}$.

Proof For fixed $x_{1:n}, \ldots, x_{n:n}$, and associated $x_1, \ldots, x_n$

$$\frac{f(x_1; \boldsymbol{\theta}) \cdots f(x_n; \boldsymbol{\theta})}{n! f(x_{1:n}; \boldsymbol{\theta}) \cdots f(x_{n:n}; \boldsymbol{\theta})} = \frac{1}{n!}$$

and zero otherwise. □

Theorem 9.1.5: Rao-Blackwell

Let $X_1, \ldots, X_n$ have joint PDF $f(x_1, \ldots, x_n; \boldsymbol{\theta})$, and let $\boldsymbol{S} = (S_1, \ldots, S_k)$ be a vector of jointly sufficient statistics for $\boldsymbol{\theta}$. If T is any unbiased estimator of $\tau(\boldsymbol{\theta})$, and if $T^* = \mathbb{E}(T|\boldsymbol{S})$, then

1. T^* is an unbiased estimator of $\tau(\boldsymbol{\theta})$,
2. T^* is a function of $\boldsymbol{S}$, and
3. $\mathrm{Var}(T^*) \le \mathrm{Var}(T)$ for every $\boldsymbol{\theta}$, and $\mathrm{Var}(T^*) < \mathrm{Var}(T)$ for some $\boldsymbol{\theta}$ unless $T^* = T$ with probability 1.

Proof Let $T = \mathscr{T}(X_1, \ldots, X_n)$, according to the factorization criterion, then

$$\mathbb{E}(T|\boldsymbol{s}) = \int \cdots \int \mathscr{T}(x_1, \ldots, x_n) f_{\boldsymbol{X}|\boldsymbol{s}}(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Thus the function $\boldsymbol{s} \mapsto \mathbb{E}(T|\boldsymbol{s})$ does not depend on $\boldsymbol{\theta}$. Therefore, $T^* = \mathbb{E}(T|\boldsymbol{S})$ is a function of $\boldsymbol{S}$, and furthermore,

$$\mathbb{E}(T^*) = \mathbb{E}_{\boldsymbol{S}}[\mathbb{E}(T|\boldsymbol{S})] = \mathbb{E}(T) = \tau(\boldsymbol{\theta}).$$

From Theorem 4.4.3,

$$\mathrm{Var}(T) = \mathrm{Var}_{\boldsymbol{S}}[\mathbb{E}(T|\boldsymbol{S})] + \mathbb{E}_{\boldsymbol{S}}[\mathrm{Var}(T|\boldsymbol{S})] \ge \mathrm{Var}_{\boldsymbol{S}}[\mathbb{E}(T|\boldsymbol{S})] = \mathrm{Var}(T^*)$$

with equality if and only if $\mathbb{E}_{\boldsymbol{S}}[\mathrm{Var}(T|\boldsymbol{S})] = 0$, which occurs if and only if $\mathrm{Var}(T|\boldsymbol{S}) = 0$ with probability 1, or equivalent $T = \mathbb{E}(T|\boldsymbol{S}) = T^*$.

9.2 Completeness

Definition 9.2.1: Completeness

A family of density function $\{f_{\boldsymbol{S}}(\boldsymbol{s}\ ,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Omega}\}$, is called **complete** if $\mathbb{E}[u(\boldsymbol{S})] = 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$ implies $u(\boldsymbol{S}) = 0$ with probability 1 for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$.

This sometimes is expressed by saying that there are no nontrivial unbiased estimators of zero. For example, two unbiased estimators $T_1 = u_1(\boldsymbol{S})$ and $T_2 = u_2(\boldsymbol{S})$, $\mathbb{E}(T_1) = \mathbb{E}(T_2) = \tau(\boldsymbol{\theta})$, then $\mathbb{E}[u_1(\boldsymbol{S}) - u_2(\boldsymbol{S})] = 0$, which implies $u_1(\boldsymbol{S}) = u_2(\boldsymbol{S})$ with probability 1, if the family of density function is complete.

Theorem 9.2.1: Lehmann-Scheffé

Let $X_1, \dots, X_n$ have joint PDF $f(x_1, \dots, x_n; \boldsymbol{\theta})$, and let $\boldsymbol{S}$ be a vector of jointly complete sufficient statistics for $\boldsymbol{\theta}$. If $T^* = \mathscr{T}^*(\boldsymbol{S})$ is a statistic that is unbiased for $\tau(\boldsymbol{\theta})$ and a function of $\boldsymbol{S}$, then T^* is a UMVUE of $\tau(\boldsymbol{\theta})$.

Proof It follows by completeness that any statistic that is a function of $\boldsymbol{S}$ and an unbiased estimator of $\tau(\boldsymbol{\theta})$ must be equal to T^* with probability 1. If T is any other statistic that is an unbiased estimator of $\tau(\boldsymbol{\theta})$, then by the Rao-Blackwell's Theorem 9.1.5 $\mathbb{E}(T|\boldsymbol{S})$ also is unbiased for $\tau(\boldsymbol{\theta})$ and a function of $\boldsymbol{S}$, so by uniqueness, $T^* = \mathbb{E}(T|\boldsymbol{S})$ with probability 1. Furthermore, $\text{Var}(T^*) \leq \text{Var}(T)$ for all $\boldsymbol{\theta}$. Thus, T^* is a UMVUE of $\tau(\boldsymbol{\theta})$. □

Definition 9.2.2: Ancillary

A statistic $\boldsymbol{T}$ is said to be **ancillary** if its distribution does not depend on $\boldsymbol{\theta}$.

Theorem 9.2.2: Basu

If $\boldsymbol{S}$ is a vector of jointly complete sufficient statistics for $\boldsymbol{\theta}$, and if $\boldsymbol{T}$ is ancillary, then $\boldsymbol{S}$ and $\boldsymbol{T}$ are stochastically independent.

Proof We will consider the discrete case. Denote by $f(\boldsymbol{t})$, $f(\boldsymbol{s}; \boldsymbol{\theta})$, and $f(\boldsymbol{t}|\boldsymbol{s})$ the PDFs of $\boldsymbol{T}$, $\boldsymbol{S}$, and the conditional PDF of $\boldsymbol{T}$ given $\boldsymbol{S} = \boldsymbol{s}$, respectively. Consider the following expanded value relative to the distribution of $\boldsymbol{S}$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{S}}[f(\boldsymbol{t}) - f(\boldsymbol{t}|\boldsymbol{S})] &= f(\boldsymbol{t}) - \sum_{\boldsymbol{s}} f(\boldsymbol{t}|\boldsymbol{S}) f(\boldsymbol{s}; \boldsymbol{\theta}) \\ &= f(\boldsymbol{t}) - \sum_{\boldsymbol{s}} f(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\theta}) \\ &= f(\boldsymbol{t}) - f(\boldsymbol{t}) = 0.\end{aligned}$$

Because $\boldsymbol{S}$ is a complete sufficient statistic, $f(\boldsymbol{t}) = f(\boldsymbol{t}|\boldsymbol{s})$, which means that $\boldsymbol{S}$ and $\boldsymbol{T}$ are stochastically independent.

The continuous case is similar. □

Definition 9.2.3: Regular Exponential Class

A density function is said to be a member of the **regular exponential class**, denoted by $\mathsf{REC}(\eta_1, \dots, \eta_k)$, if it can be expressed in the form

$$f(x; \boldsymbol{\theta}) = c(\boldsymbol{\theta}) h(x) \exp\left[\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta})_j(x)\right], \quad x \in A \tag{9.2.1}$$

and zero otherwise, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ is a vector of k unknown parameters, if the parameter space has the form

$$\boldsymbol{\Omega} = \{\boldsymbol{\theta} | a_i \le \theta_i \le b_i,\ i = 1, \dots, k\}$$

(note that $a_i = -\infty$ and $b_i = \infty$ are permissible values), and if it satisfies the following regularity conditions:

1. The set $A = \{x : f(x, \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.
2. The function $\eta_j(\boldsymbol{\theta})$ are nontrivial, functionally independent, continuous functions of the θ_i.

3a. For a continuous random variable, the derivative $\mathscr{T}_j'(x)$ are linearly independent continuous functions of x over A.

3b. For a discrete random variable, the $\mathscr{T}_j(x)$ are nontrivial functions of x on A, and none is a linear function of the others.

Theorem 9.2.3

If $X_1, \dots, X_n$ is a random sample from a member of $\mathsf{REC}(\eta_1, \dots, \eta_k)$, then the statistics

$$S_j = \sum_{i=1}^{n} \mathscr{T}_j(X_i), \quad j = 1, \dots, k$$

are a minimal set of complete sufficient statistics for $\boldsymbol{\theta}$.

If we call an estimator whose variance achieves the CRLB a CRLB estimator, then the following theorems can be stated.

Theorem 9.2.4

If a CRLB estimator T exists for $\tau(\theta)$, then a single sufficient statistic exists, and T is a function of the sufficient statistic. Conversely, if a single sufficient statistic exists and the CRLB exists, then a CRLB estimator exists for some $\tau(\theta)$.

Theorem 9.2.5

If the CRLB exists, then a CRLB estimator will exist for some function $\tau(\theta)$ if and only if the density function is a member of the REC. Furthermore, the CRLB estimator of $\tau(\theta)$ will be $\tau(\hat{\theta})$, where $\hat{\theta}$ is the MLE of θ.

Example 9.2.1

Consider a Bernoulli distribution, $X \sim \mathsf{BIN}(1, p)$. It follows that

$$f(x;\, p) = p^x(1-p)^{1-x} = (1-p)\exp\left(x \ln \frac{p}{1-p}\right), \quad x \in A = \{0, 1\}$$

which is $\mathsf{REC}(\eta_1)$ with $\eta_1(p) = \ln[p/(1-p)]$ and $\mathscr{T}_1(x) = x$. Therefore, $S_1 = \sum_{i=1}^n X_i$ is a complete sufficient statistic for p.

Let $T = \bar{X} = S_1/n$, then $\mathbb{E}(T) = p$, so T is unbiased estimator of p, according to Lehmann-Scheffé's Theorem 9.2.1, T is a UMVUE of p.

Since $\ln f(x;\, p) = x \ln p + (1-x)\ln(1-p)$,

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[\left(\frac{\partial}{\partial p} \ln f(X;\, p)\right)^2\right] = \mathbb{E}\left[\left(\frac{X}{p} - \frac{1-X}{1-p}\right)^2\right] \\ &= \mathbb{E}\left[\left(\frac{X-p}{p(1-p)}\right)^2\right] = \frac{\mathrm{Var}(X-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)} \end{aligned}$$

the CRLB of p is $1/[nI(\theta)] = p(1-p)/n$. And

$$\mathrm{Var}(T) = \frac{1}{n^2}\mathrm{Var}(S_1) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{1}{n^2} n \,\mathrm{Var}(X) = \frac{1}{n} p(1-p),$$

so $T = S_1/n$ is CRLB estimator of p.

Example 9.2.2

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$f(x;\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right]$$

which is $\mathsf{REC}(\boldsymbol{\eta})$ with $\boldsymbol{\eta}(\mu, \sigma) = [\mu/\sigma^2,\ -1/(2\sigma^2)]^\top$, and $\mathscr{T}(x) = [x,\ x^2]^\top$. Therefore, $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are a complete sufficient statistics for μ and σ.

Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right),$$

then by Theorem 7.1.5, we have $\mathbb{E}(\bar{X}) = \mu$ and $\mathbb{E}(S^2) = \sigma^2$. According to Lehmann-Scheffé's Theorem 9.2.1, we have the following UMVUEs:

- The UMVUE of μ is $\bar{X}$.
- The UMVUE of σ^2 is S^2.
- The UMVUE of μ^2 is $\bar{X}^2 - S^2/n$, since $\mathbb{E}(\bar{X}^2 - S^2/n) = \mathbb{E}(\bar{X}^2) - \mathbb{E}(S^2)/n = \mu^2$.
- The UMVUE of σ^r with $r > 1-n$ is $k_{n-1,r}S^r$, where

$$k_{m,r} = \frac{n^{r/2}\Gamma\left(\frac{m}{2}\right)}{2^{r/2}\Gamma\left(\frac{m+r}{2}\right)}.$$

  This is caused by the fact that $\sqrt{n-1}S/\sigma \sim \chi(n-1)$, and

$$\mathbb{E}(S^r) = \frac{\sigma^r}{(n-1)^{r/2}}\mathbb{E}\left[\left(\frac{\sqrt{n-1}S}{\sigma}\right)^r\right] = \frac{\sigma^r}{(n-1)^{r/2}} \cdot 2^{r/2}\frac{\Gamma\left(\frac{n-1+r}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

- The UMVUE of μ/σ is $k_{n-1,-1}\bar{X}/S$, if $n > 2$. This is caused by the fact that $\sqrt{n}\bar{X}/S \sim t(n-1, \sqrt{n}\mu/\sigma)$, and

$$\mathbb{E}(\bar{X}/S) = \frac{\mu}{\sigma}\sqrt{\frac{n-1}{2}}\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

- Suppose that ϑ satisfies $\mathbb{P}(X_1 \le \vartheta) = p$ with a fixed $p \in (0,1)$. Then $\vartheta = \mu + \sigma\Phi^{-1}(p)$ and its UMVUE is $\bar{X} + k_{n-1,1}S\Phi^{-1}(p)$.
- Let c be a fixed constant and $\vartheta = \mathbb{P}(X_1 \le c) = \Phi\left(\frac{c-\mu}{\sigma}\right)$. Since $I_{(-\infty,c)}(X_1)$ is an unbiased estimator of ϑ, the UMVUE of ϑ is $\mathbb{E}[I_{(-\infty,c)}(X_1)|\boldsymbol{S}] = \mathbb{P}(X_1 \le c|\boldsymbol{S})$. By Basu's Theorem 9.2.2, $Z = (X_1 - \bar{X})/S$ is independent of $\boldsymbol{S} = (\bar{X}, S^2)$. Since

$$\mathbb{P}(X_1 \le c|\boldsymbol{S} = (\bar{x}, s^2)) = \mathbb{P}\left(Z \le \frac{c-\bar{X}}{S}\middle|\boldsymbol{S} = (\bar{x}, s^2)\right) = \mathbb{P}\left(Z \le \frac{c-\bar{x}}{s}\right),$$

  the UMVUE of ϑ is

$$\mathbb{P}(X_1 \le c|\boldsymbol{S}) = \int_{-(n-1)/\sqrt{n}}^{(c-\bar{X})/S} f(z; n)\,\mathrm{d}z,$$

  with f given by (7.3.1).

- Suppose that we would like to estimate $\vartheta = \frac{1}{\sigma}\Phi'\left(\frac{c-\mu}{\sigma}\right) = f_X(c)$. Since the conditional PDF of X_1 given $\bar{X} = \bar{x}$ and $S^2 = s^2$ is

$$f_{X_1|\boldsymbol{S}=(\bar{x},s^2)}(c) = \left[\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(X_1 \le x|\boldsymbol{S} = (\bar{x}, s^2))\right]_{x=c} = \frac{1}{s}f\left(\frac{c-\bar{x}}{s}; n\right).$$

Let $f_{\boldsymbol{S}}$ be the joint PDF of $\boldsymbol{S} = (\bar{X}, S^2)$. By Law of Total Expectation, we have

$$\begin{aligned}\vartheta &= \mathbb{E}\left[I_{\{c\}}(X_1)\right] = \mathbb{E}_{\boldsymbol{S}}\left[\mathbb{E}\left(I_{\{c\}}(X_1)|\boldsymbol{S}\right)\right] \\ &= \iint \left[\int I_{\{c\}}(x) f_{X_1|\boldsymbol{S}=(\bar{x},s^2)}(x)\,\mathrm{d}x\right] \mathrm{d}\boldsymbol{s} \\ &= \iint f_{X_1|\boldsymbol{S}=(\bar{x},s^2)}(c)\,\mathrm{d}\boldsymbol{s} = \mathbb{E}\left[\frac{1}{S}f\left(\frac{c-\bar{X}}{S}; n\right)\right].\end{aligned}$$

Hence the UMVUE of ϑ is $\frac{1}{S}f\left(\frac{c-\bar{X}}{S}; n\right)$.

Definition 9.2.4: Range-Dependent Exponential Class

A density function is said to be a member of the **range-dependent exponential class**, denoted by RDEC$(\eta_1, \ldots, \eta_k)$, if it satisfies regularity conditions 2 and 3a or 3b of Definition 9.2.3, for $j = 3, \ldots, k$, and if it has the form

$$f(x; \boldsymbol{\theta}) = c(\boldsymbol{\theta})h(x)\exp\left[\sum_{j=3}^{k}\eta_j(\theta_3, \ldots, \theta_k)\mathscr{T}_j(x)\right], \quad x \in A \tag{9.2.2}$$

where $A = \{\eta_1(\theta_1, \theta_2) < x < \eta_2(\theta_1, \theta_2)\}$ and $\boldsymbol{\theta} \in \boldsymbol{\Omega}$.

We will include the following special cases:

1. The one-parameter case, where $f(x;\ \theta) = c(\theta)h(x)$ with $A = \{x|\eta_1(\theta) < x < \eta_2(\theta)\}$.
2. The two-parameter case, where $f(x;\ \theta_1, \theta_2) = c(\theta_1, \theta_2)h(x)$ with $A = \{x|\eta_1(\theta_1, \theta_2) < x < \eta_2(\theta_1, \theta_2)\}$.

Theorem 9.2.6

Let $X_1, \ldots, X_n$ be a random sample from a member of the RDEC$(\eta_1, \ldots, \eta_k)$.

1. In the one-parameter case, $S_1 = X_{1:n}$ and $S_2 = X_{n:n}$ are jointly sufficient for θ.

   (a) $T = \min[\eta_1^{-1}(X_{1:n}), \eta_2^{-1}(X_{n:n})]$, if $\eta_1(\theta)$ is increasing and $\eta_2(\theta)$ is decreasing,

   (b) $T = \max[\eta_1^{-1}(X_{1:n}), \eta_2^{-1}(X_{n:n})]$ if $\eta_1(\theta)$ is decreasing and $\eta_2(\theta)$ is increasing,

   is a single sufficient statistic for θ.

2. In the two-parameter case, $S_1 = X_{1:n}$ and $S_2 = X_{n:n}$ are jointly sufficient for $\boldsymbol{\theta} =$

(θ_1, θ_2).

3. If $k > 2$, then $S_1 = X_{1:n}$, $S_2 = X_{n:n}$ and $S_3, \ldots, S_k$ where $S_j = \sum_{i=1}^n \mathscr{T}_j(X_i)$ are jointly sufficient for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

Theorem 9.2.7

Suppose that $X_1, \ldots, X_n$ be a random sample from a member of the $\mathsf{RDEC}(\eta_1, \ldots, \eta_k)$.

1. In the one-parameter case, $S_1 = X_{1:n}$ and $S_2 = X_{n:n}$ are jointly sufficient for θ.

   (a) $S_2 = X_{n:n}$, if $\eta_1(\theta)$ does not depend on θ,

   (b) $S_1 = X_{1:n}$, if $\eta_2(\theta)$ does not depend on θ,

   is a single sufficient statistic for θ.

2. If $k > 2$, and if

   (a) the lower limit is constant, say $\eta_1(\theta) = a$, then $S_2 = X_{n:n}$

   (b) the upper limit is constant, say $\eta_2(\theta) = b$, then $S_1 = X_{1:n}$

   and $S_3, \ldots, S_k$ where $S_j = \sum_{i=1}^n \mathscr{T}_j(X_i)$ are jointly sufficient for θ and θ_j; $j = 3, \ldots, k$.

Example 9.2.3

Consider a random sample of size n from a uniform distribution, $X_i \sim \mathsf{UNIF}(a, b)$:

$$f(x;\, a, b) = \frac{1}{b-a} I_{(a,b)}(x).$$

If $X_1, \ldots, X_n$ is a random sample, then it follows from Theorem 9.2.6 that $X_{1:n}$ and $X_{n:n}$ are **jointly sufficient** for (a, b). We also can verify the sufficiency by Fisher-Neyman factorization criterion (Theorem 9.1.1): the joint PDF of $X_1, \ldots, X_n$ is

$$\begin{aligned} f(x_1, \ldots, x_n;\, a, b) &= \left(\frac{1}{b-a}\right)^n \prod_{i=1}^n I_{(a,b)}(x_i) \\ &= \left(\frac{1}{b-a}\right)^n I_{(a,b)}(x_{1:n}) I_{(a,b)}(x_{n:n}) \\ &= g(x_{1:n}, x_{n:n};\, a, b). \end{aligned}$$

According to Theorem 5.5.2, the PDFs and joint PDF of $X_{1:n}$ and $X_{n:n}$ are given by

$$\begin{aligned}
f_{X_{1:n}}(x) &= n\frac{1}{b-a}\left(1-\frac{x-a}{b-a}\right)^{n-1} I_{(a,b)}(x), \\
f_{X_{n:n}}(x) &= n\left(\frac{x-a}{b-a}\right)^{n-1}\frac{1}{b-a}I_{(a,b)}(x), \\
f_{(X_{1:n},X_{n,n})}(x,y) &= \frac{n(n-1)}{(b-a)^2}\left(\frac{y-x}{b-a}\right)^{n-2} I_{(a,b)}(x)I_{(x,b)}(y).
\end{aligned}$$

To verify **completeness**, assume that $\mathbb{E}[u(X_{1:n},X_{n:n})]=0$ for all $a<b$, which means that

$$\begin{aligned}
\mathbb{E}[u(X_{1:n},X_{n:n})] &= \int_a^b\int_a^y u(x,y)f_{(X_{1:n},X_{n,n})}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{n(n-1)}{(b-a)^n}\int_a^b\int_a^y u(x,y)(y-x)^{n-2}\,\mathrm{d}x\,\mathrm{d}y = 0.
\end{aligned}$$

Then

$$0 = \frac{\partial^2}{\partial a\partial b}\int_a^b\int_a^y u(x,y)(y-x)^{n-2}\,\mathrm{d}x\,\mathrm{d}y = -(b-a)^{n-2}u(a,b),$$

which implies $u(X_{1:n},X_{n:n})=0$, and thus the jointly sufficient statistics $X_{1:n}$ and $X_{n:n}$ are also complete.

Since

$$\begin{aligned}
\mathbb{E}(X_{1:n}) &= \int_a^b xf_{X_{1:n}}(x)\,\mathrm{d}x = \frac{na+b}{n+1}, \\
\mathbb{E}(X_{n:n}) &= \int_a^b xf_{X_{n:n}}(x)\,\mathrm{d}x = \frac{a+nb}{n+1},
\end{aligned}$$

if we let

$$\hat{a} = \frac{nX_{1:n}-X_{n:n}}{n-1}, \quad \hat{b} = \frac{nX_{n:n}-X_{1:n}}{n-1},$$

then $\mathbb{E}(\hat{a})=a$ and $\mathbb{E}(\hat{b})=a$. Therefore, $\hat{a}$ is a UMVUE of a, and $\hat{b}$ is a UMVUE of b.

The CDFs of $\hat{a}$ and $\hat{b}$ are given by

$$
\begin{aligned}
F_{\hat{a}}(z) &= \mathbb{P}(\hat{a} \le z) = \mathbb{P}\left(\frac{nX_{1:n} - X_{n:n}}{n-1} \le z\right) = \iint\limits_{nx-y\le(n-1)z} f_{(X_{1:n},X_{n,n})}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left[1 - \left(1-\frac{1}{n}\right)^{n-1}\left(\frac{b-z}{b-a}\right)^{n}\right] I_{(\xi,b)}(z) + (n-1)^{n-1}\left(\frac{a-z}{b-a}\right)^{n} I_{(\xi,a)}(z) + I_{[b,\infty)}(z), \\
F_{\hat{b}}(z) &= \mathbb{P}(\hat{b} \le z) = \mathbb{P}\left(\frac{nX_{n:n} - X_{1:n}}{n-1} \le z\right) = \iint\limits_{ny-x\le(n-1)z} f_{(X_{1:n},X_{n,n})}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \left(1-\frac{1}{n}\right)^{n-1}\left(\frac{z-a}{b-a}\right)^{n} I_{(a,\eta)}(z) - (n-1)^{n-1}\left(\frac{z-b}{b-a}\right)^{n} I_{(b,\eta)}(z) + I_{[\eta,\infty)}(z),
\end{aligned}
$$

where

$$
\xi = \frac{na-b}{n-1}, \quad \eta = \frac{nb-a}{n-1},
$$

and note $\xi < a < b < \eta$. Then the PDFs of $\hat{a}$ and $\hat{b}$ are given by

$$
\begin{aligned}
f_{\hat{a}}(z) &= \frac{n}{b-a}\left[\left(1-\frac{1}{n}\right)^{n-1}\left(\frac{b-z}{b-a}\right)^{n-1} I_{(\xi,b)}(z) - (n-1)^{n-1}\left(\frac{a-z}{b-a}\right)^{n-1} I_{(\xi,a)}(z)\right], \\
f_{\hat{b}}(z) &= \frac{n}{b-a}\left[\left(1-\frac{1}{n}\right)^{n-1}\left(\frac{z-a}{b-a}\right)^{n-1} I_{(a,\eta)}(z) - (n-1)^{n-1}\left(\frac{z-b}{b-a}\right)^{n-1} I_{(b,\eta)}(z)\right].
\end{aligned}
$$

Let

$$
\begin{aligned}
R &= X_{n:n} - X_{1:n}, \\
P &= nX_{1:n} - X_{n:n} - (n-1)a = (n-1)(\hat{a} - a), \\
Q &= nX_{n:n} - X_{1:n} - (n-1)b = (n-1)(\hat{b} - b),
\end{aligned}
$$

then

$$
\begin{cases}
X_{1:n} = a + \dfrac{P+R}{n-1}, \\
X_{n:n} = a + \dfrac{P+nR}{n-1},
\end{cases}
\qquad
\begin{cases}
X_{1:n} = b + \dfrac{Q-nR}{n-1}, \\
X_{n:n} = b + \dfrac{Q-R}{n-1},
\end{cases}
$$

and

$$
\left|\frac{\partial(X_{1:n}, X_{n,n})}{\partial(P,R)}\right| = \left|\frac{\partial(X_{1:n}, X_{n,n})}{\partial(Q,R)}\right| = \frac{1}{n-1}.
$$

Thus, the joint PDFs and the PDF of R are given by

$$\begin{aligned} f_{(P,R)}(p,r) &= \frac{nr^{n-2}}{(b-a)^n} I_{(0,b-a)}(r) I_{(-r,-nr+(n-1)(b-a))}(p), \\ f_{(Q,R)}(q,r) &= \frac{nr^{n-2}}{(b-a)^n} I_{(0,b-a)}(r) I_{(nr-(n-1)(b-a),r)}(q), \\ f_R(r) &= \frac{n(n-1)r^{n-2}}{(b-a)^n}(b-a-r) I_{(0,b-a)}(r). \end{aligned}$$

The PDF of $U = P/R$ and the PDF of $V = Q/R$ are given by

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} |r| f_{(P,R)}(ur,r)\,\mathrm{d}r = I_{(-1,\infty)}(u) \int_0^{\frac{(n-1)(b-a)}{n+u}} \frac{nr^{n-1}}{(b-a)^n}\,\mathrm{d}r = \left(\frac{n-1}{n+u}\right)^n I_{(-1,\infty)}(u), \\ f_V(v) &= \int_{-\infty}^{\infty} |r| f_{(Q,R)}(vr,r)\,\mathrm{d}r = I_{(-\infty,1)}(v) \int_0^{\frac{(n-1)(b-a)}{n-v}} \frac{nr^{n-1}}{(b-a)^n}\,\mathrm{d}r = \left(\frac{n-1}{n-v}\right)^n I_{(-\infty,1)}(v). \end{aligned}$$

By the PDFs of U and V, if we have λ_1 and λ_2 such that

$$\begin{aligned} \mathbb{P}(\lambda_1 < U < \lambda_2) = 1-\alpha \quad &\Rightarrow \quad \hat{a} - \frac{\lambda_2 R}{n-1} < a < \hat{a} - \frac{\lambda_1 R}{n-1}, \\ \mathbb{P}(\lambda_1 < V < \lambda_2) = 1-\alpha \quad &\Rightarrow \quad \hat{b} - \frac{\lambda_2 R}{n-1} < b < \hat{b} - \frac{\lambda_1 R}{n-1}. \end{aligned}$$

Example 9.2.4

Consider a two-parameter exponential distribution on (a,∞) with scale parameter $\theta > 0$, i.e. $X \sim \mathsf{EXP}(\theta, a)$:

$$\begin{aligned} f(x;\theta,a) &= \frac{1}{\theta}\exp\left(-\frac{x-a}{\theta}\right) I_{(a,\infty)}(x) \\ &= \frac{1}{\theta}\exp\frac{a}{\theta}\exp\left(-\frac{x}{\theta}\right) I_{(a,\infty)}(x). \end{aligned}$$

This distribution has the following properties:

- If $Y = kX + c$, then $Y \sim \mathsf{EXP}(k\theta, ka + c)$.
- If $Y = 2(X-a)/\theta$, then $f_Y(y) = \frac{1}{2}\exp(-y/2)$, thus $Y \sim \mathsf{EXP}(2) = \chi^2(2)$.
- If $X_i \sim \mathsf{EXP}(\theta_i, a);\ i = 1,\ldots,n$, then

$$\min(X_1,\ldots,X_n) \sim \mathsf{EXP}\left(\frac{1}{\sum_{i=1}^n \theta_i^{-1}}, a\right).$$

If $X_1,\ldots,X_n$ is a random sample from $\mathsf{EXP}(\theta,a)$, then it follows from Theorem 9.2.6 that $X_{1:n}$ and $\sum_{i=1}^n X_i$ are **jointly sufficient** for (θ,a). We also can verify the sufficiency by Fisher-Neyman factorization criterion (Theorem 9.1.1): the joint PDF of $X_1, \ldots, X_n$ is

$$\begin{aligned} f(x_1, \ldots, x_n; \theta, a) &= \frac{1}{\theta^n} \exp \frac{na}{\theta} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\theta}\right) \prod_{i=1}^{n} I_{(a,\infty)}(x_i) \\ &= \frac{1}{\theta^n} \exp \frac{na}{\theta} \exp\left(-\sum_{i=1}^{n} \frac{x_i}{\theta}\right) I_{(a,\infty)}(x_{1:n}) \\ &= g\left(x_{1:n}, \sum_{i=1}^{n} x_i; \theta, a\right). \end{aligned}$$

According to Theorem 5.5.2 and the properties, we have $X_{1:n} \sim \mathrm{EXP}(\theta/n, a)$ and $2n(\bar{X} - X_{1:n})/\theta \sim \chi^2(2n-2)$. Thus, $\mathbb{E}(X_{1:n}) = a + \theta/n$ and $\mathbb{E}(\bar{X} - X_{1:n}) = (n-1)\theta/n$. Therefore, the UMVUE of θ is $n(\bar{X} - X_{1:n})/(n-1)$, and the UMVUE of a is $X_{1:n} - (\bar{X} - X_{1:n})/(n-1)$.

Because, for each fixed value of θ, $X_{1:n}$ is complete ($\mathbb{E}[u(X_{1:n})] = 0$ implies $u(X_{1:n}) = 0$ for all $a \in \mathbb{R}$) and sufficient (Theorem 9.2.7) for a, from Basu's Theorem 9.2.2, $X_{1:n}$ and $\bar{X} - X_{1:n}$ are stochastically independent[a]. Since $2n(X_{1:n} - a)/\theta \sim \chi^2(2)$, we have

$$(n-1)\frac{X_{1:n} - a}{\bar{X} - X_{1:n}} \sim F(2, 2n-2), \quad \frac{X_{1:n} - a}{\bar{X} - a} \sim \mathsf{BETA}(1, n-1).$$

[a] This can also be verified by the covariance of $X_{1:n}$ and $\bar{X} - X_{1:n}$. Rahman & Pearson (2001) shows that $\mathrm{Cov}(X_{1:n}, X_{i:n}) = \theta^2/n^2$ for $i = 1, \ldots, n$. Therefore, $\mathrm{Cov}(X_{1:n}, \bar{X} - X_{1:n}) = 0$.

Chapter 10

Interval Estimation

10.1 Confidence Intervals

Definition 10.1.1: Confidence Interval

An interval $(\mathscr{L}(x_1, \ldots, x_n), \mathscr{U}(x_1, \ldots, x_n))$ is called a $100\gamma\%$ **confidence interval** for θ if

$$\mathbb{P}[\mathscr{L}(X_1, \ldots, X_n) < \theta < \mathscr{U}(X_1, \ldots, X_n)] = \gamma \tag{10.1.1}$$

where $0 < \gamma < 1$ called the **confidence coefficient** or **confidence level**. The observed values $\mathscr{L}(x_1, \ldots, x_n)$ and $\mathscr{U}(x_1, \ldots, x_n)$ are called **lower** and **upper confidence limits**, respectively.

Generally speaking, for a prescribed confidence level, we want to use a method that produces an interval with some optimal property such as minimal length. Actually, the length, $\mathscr{U}(X_1, \ldots, X_n) - \mathscr{L}(X_1, \ldots, X_n)$ of the corresponding random interval generally will be a random variable, so a criterion such as **minimum expected length** might be more appropriate. For some problems, the **equal tailed** choice will provide the minimum expected length, but for others it will not.

Definition 10.1.2: One-Sided Confidence Limits

1. If

$$\mathbb{P}[\mathscr{L}(X_1, \ldots, X_n) < \theta] = \gamma \tag{10.1.2}$$

then $\mathscr{L}(x_1, \ldots, x_n)$ is called a **one-sided lower** $100\gamma\%$ **confidence limits** for θ.

2. If

$$\mathbb{P}[\theta < \mathscr{U}(X_1, \ldots, X_n)] = \gamma \tag{10.1.3}$$

then $\mathscr{U}(x_1, \ldots, x_n)$ is called a **one-sided upper** $100\gamma\%$ **confidence limits** for θ.

10.2 Pivotal Quantity Method

Definition 10.2.1: Pivotal Quantity

If $Q = \mathscr{Q}(X_1, \ldots, X_n; \theta)$ is a random variable that is a function only of $X_1, \ldots, X_n$ and θ, then Q is called a **pivotal quantity** if its distribution does not depend on θ or any other unknown parameters.

If Q is a pivotal quantity for a parameter θ and if percentiles of Q, say q_1 and q_2, are available such that

$$\mathbb{P}[q_1 < \mathscr{Q}(X_1, \ldots, X_n; \theta) < q_2] = \gamma, \tag{10.2.1}$$

then for an observed sample, $x_1, \ldots, x_n$, a $100\gamma\%$ confidence region for θ is

$$\{\theta \in \Omega | q_1 < \mathscr{Q}(x_1, \ldots, x_n; \theta) < q_2\}. \tag{10.2.2}$$

Definition 10.2.2: Location & Scale Parameter

Let $f_0(z)$ be a PDF that is free of unknown parameters, then

1. θ is a **location parameter**, if the PDF has the form $f(x; \theta) = f_0(x - \theta)$.
2. θ is a **scale parameter**, if the PDF has the form $f(x; \theta) = (1/\theta)f_0(x/\theta)$.
3. θ_1, θ_2 are **location-scale parameters** if the PDF has the form

$$f(x; \theta_1, \theta_2) = (1/\theta_2)f_0[(x - \theta_1)/\theta_2].$$

Theorem 10.2.1

Let $X_1, \ldots, X_n$ be a random sample from a distribution with PDF $f(x; \theta)$ for $\theta \in \Omega$, and assume that an MLE $\hat{\theta}$ exists.

1. If θ is a location parameter, then $Q = \hat{\theta} - \theta$ is a pivotal quantity.
2. If θ is a scale parameter, then $Q = \hat{\theta}/\theta$ is a pivotal quantity.

Theorem 10.2.2

Let $X_1, \ldots, X_n$ be a random sample of sizen from a distribution with PDF of the form:

$$f(x; \theta_1, \theta_2, \kappa) = \frac{1}{\theta_2} f_0\left(\frac{x - \theta_1}{\theta_2}; \kappa\right),$$

where $-\infty < \theta_1 < \infty$ and $\theta_2 > 0$, and $f_0(z; \kappa)$ is a PDF that depends on κ but not on θ_1 and θ_2. If there exist MLEs $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\kappa}$, then the distribution of $(\hat{\theta}_1 - \theta_1)/\hat{\theta}_2$, $\hat{\theta}_2/\theta_2$ and $\hat{\kappa}$ do not depend on θ_1 and θ_2.

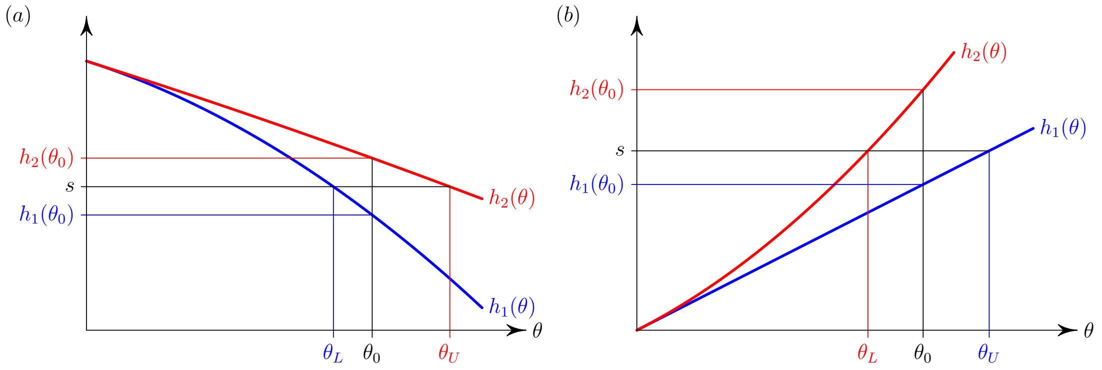

Figure 10.1: A confidence interval based on the general method. $h_1(\theta)$ and $h_2(\theta)$ are monotonic decreasing (figure (a)) or increasing (figure (b)) functions of θ.

10.3 General Method

If a pivotal quantity is not available, then it is still possible to determine a confidence region for a parameter θ if a statistic exists with a distribution that depends on θ but not on any other unknown nuisance parameters. Specifically, let $X_1, \ldots, X_n$ have joint PDF $f(x_1, \ldots, x_n; \theta)$, and $S = \mathscr{S}(X_1, \ldots, X_n) \sim g(s; \theta)$. Preferably S will be sufficient for θ, or possibly some reasonable estimator such as an MLE, but this is not required.

Now, for each possible value of θ, assume that we can find values $h_1(\theta)$ and $h_2(\theta)$ such that

$$\mathbb{P}[h_1(\theta) < S < h_2(\theta)] = 1 - \alpha. \tag{10.3.1}$$

If we observe $S = s$, then the set of values $\theta \in \Omega$ that satisfy $h_1(\theta) < s < h_2(\theta)$ form a $100(1-\alpha)\%$ confidence region. In other words, if θ_0 is the true value of θ, then θ_0 will be in the confidence region if and only if $h_1(\theta_0) < s < h_2(\theta_0)$, which has $100(1-\alpha)\%$ confidence level because equation (10.3.1) holds with $\theta = \theta_0$ in this case. Quite often $h_1(\theta)$ and $h_2(\theta)$ will be monotonic decreasing (or increasing) functions of θ, and the resulting confidence region will be an interval (see figure 10.1).

Theorem 10.3.1

Let the statistic S be continuous with CDF $G(s; \theta)$, and suppose that $h_1(\theta)$ and $h_2(\theta)$ are functions that satisfy

$$\mathbb{P}[S \leq h_1(\theta); \theta] = \alpha_1 \quad \Leftrightarrow \quad G(h_1(\theta); \theta) = \alpha_1, \tag{10.3.2}$$
$$\mathbb{P}[S \geq h_2(\theta); \theta] = \alpha_2 \quad \Leftrightarrow \quad G(h_2(\theta); \theta) = 1 - \alpha_2, \tag{10.3.3}$$

for each $\theta \in \Omega$, where $\alpha_1, \alpha_2 \in (0, 1)$. Let s be an observed value of S.

- If $G(s; \theta)$ is a increasing function of θ, then $h_1(\theta)$ and $h_2(\theta)$ are decreasing, and:
  1. A one-sided lower $100(1-\alpha_1)\%$ confidence limit is a solution of $h_1(\theta_L) = s$ or $G(s; \theta_L) = \alpha_1$;

2. A one-sided upper $100(1-\alpha_2)\%$ confidence limit is a solution of $h_2(\theta_U) = s$ or $G(s;\theta_U) = 1-\alpha_2$;

- If $G(s;\theta)$ is a decreasing function of θ, then $h_1(\theta)$ and $h_2(\theta)$ are increasing, and:

  1. A one-sided lower $100(1-\alpha_2)\%$ confidence limit is a solution of $h_2(\theta_L) = s$ or $G(s;\theta_L) = 1-\alpha_2$;
  2. A one-sided upper $100(1-\alpha_1)\%$ confidence limit is a solution of $h_1(\theta_U) = s$ or $G(s;\theta_U) = \alpha_1$;

- If $\alpha_1 + \alpha_2 = \alpha \in (0,1)$, then (θ_L, θ_U) is a $100(1-\alpha)\%$ confidence interval for θ.

Definition 10.3.1: Conservative Confidence Interval

An observed confidence interval (θ_L, θ_U) is called a **conservative** $100(1-\alpha)\%$ confidence interval for θ if the corresponding random interval contains the true value of θ with probability **at least** $1-\alpha$.

10.4 Bayesian Interval Estimation

In any event, suppose that a prior density $p(\theta)$ exists or is introduced into the problem and $f(x;\theta)$ is interpreted as a conditional PDF, $f(x|\theta)$. Consider again the posterior density of θ given the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$,

$$f_{\theta|\boldsymbol{x}}(\theta) = \frac{f(x_1,\ldots,x_n|\theta)p(\theta)}{\displaystyle\int f(x_1,\ldots,x_n|\theta)p(\theta)\,\mathrm{d}\theta}. \tag{10.4.1}$$

The prior density $p(\theta)$ can be interpreted as specifying an initial probability distribution for the possible values of θ, and in this context $f_{\theta|\boldsymbol{x}}(\theta)$ would represent a revised distribution adjusted by the observed random sample. For a particular $1-\alpha$ level, a Bayesian confidence interval for θ is given by (θ_L, θ_U) where θ_L and θ_U satisfy

$$\int_{\theta_L}^{\theta_U} f_{\theta|\boldsymbol{x}}(\theta)\,\mathrm{d}\theta = 1-\alpha. \tag{10.4.2}$$

Chapter 11

Tests of Hypotheses

11.1 Introduction

Suppose that we partition the parameter space Ω into two disjoint sets Ω_0 and $\Omega_1 = \Omega - \Omega_0$ and that we wish to test

$$H_0 : \theta \in \Omega_0 \quad \text{versus} \quad H_1 : \theta \in \Omega_1. \tag{11.1.1}$$

We call H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

> **Definition 11.1.1**
>
> If $X \sim f(x;\theta)$, a statistical hypothesis is a statement about the distribution of X. If the hypothesis completely specifies $f(x;\theta)$, then it is referred to as a **simple** hypothesis; otherwise it is called **composite**.

For example, a hypothesis of the form $\theta = \theta_0$ is called a **simple** hypothesis, and a hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite** hypothesis.

In order to test which of the two hypotheses, **null hypothesis** H_0 or **alternative hypothesis** H_1, is true, we shall set up a rule based on $x_1, x_2, \ldots, x_n$ (the observed values of a random sample of size n). The rule leads to a decision to accept or reject H_0; hence, it is necessary to partition the sample space into two parts – say, $\mathcal{R}$ and $\mathcal{R}^c$ – so that

- if $(x_1, x_2, \ldots, x_n) \in \mathcal{R}$, reject H_0;
- if $(x_1, x_2, \ldots, x_n) \in \mathcal{R}^c$, accept (do not reject) H_0.

The rejection region $\mathcal{R}$ for H_0 is called the **critical region** for the test.

	Retain H_0	Reject H_0
H_0 is true	√	Type I error
H_1 is true	Type II error	√

Table 11.1: Summary of outcomes of hypothesis testing.

There are two types errors we can make. Rejecting H_0 when H_0 is true is called **Type I error**. Retaining H_0 when H_1 is true is called **Type II error**. The possible outcomes for hypothesis testing are summarized in Table 11.1. We will adopt the following notations for these error probabilities:

$$\alpha = \mathbb{P}(\text{Type I error}) = \mathbb{P}\Big((X_1, \ldots, X_n) \in \mathcal{R} \mid H_0\Big), \tag{11.1.2}$$

$$\beta = \mathbb{P}(\text{Type II error}) = \mathbb{P}\Big((X_1, \ldots, X_n) \in \mathcal{R}^c \mid H_1\Big). \tag{11.1.3}$$

Definition 11.1.2: Power Function

The **power function** of a test with rejection region $\mathcal{R}$ is defined by

$$\pi(\theta) = \mathbb{P}\Big((X_1, \ldots, X_n) \in \mathcal{R}; \theta\Big). \tag{11.1.4}$$

Definition 11.1.3: Significance Level & Size

- For a simple null hypothesis, $H_0 : \theta = \theta_0$, the probability of rejecting a true H_0,

  $$\alpha = \mathbb{P}(\text{Type I error}) = \pi(\theta_0) \tag{11.1.5}$$

  is referred to as the **significance level** of the test.

- For a composite null hypothesis, $H_0 : \theta \in \Omega_0$, the **size** of the test (or size of the critical region) is the maximum probability of rejecting H_0 when H_0 is true:

  $$\alpha = \sup_{\theta \in \Omega_0} \pi(\theta). \tag{11.1.6}$$

For simple hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1,$$

we have

$$\alpha = \pi(\theta_0), \quad \beta = 1 - \pi(\theta_1). \tag{11.1.7}$$

For composite hypotheses:

$$H_0 : \theta \in \Omega_0 \quad \text{versus} \quad H_1 : \theta \in \Omega_1,$$

the size of the test (or critical region) is

$$\alpha = \sup_{\theta \in \Omega_0} \pi(\theta), \tag{11.1.8}$$

and if the true value θ falls in Ω_1, then

$$\beta = 1 - \pi(\theta), \tag{11.1.9}$$

where we note that $\mathbb{P}(\text{Type II error})$ depends on θ.

The value of the power function is always the area under the PDF of the test statistic and over the critical region, giving $\mathbb{P}(\text{Type I error})$ for values of θ in the null hypothesis and $1 - \mathbb{P}(\text{Type II error})$ for values of θ in the alternative hypothesis. This is illustrated for a test of means in Figure 11.1.

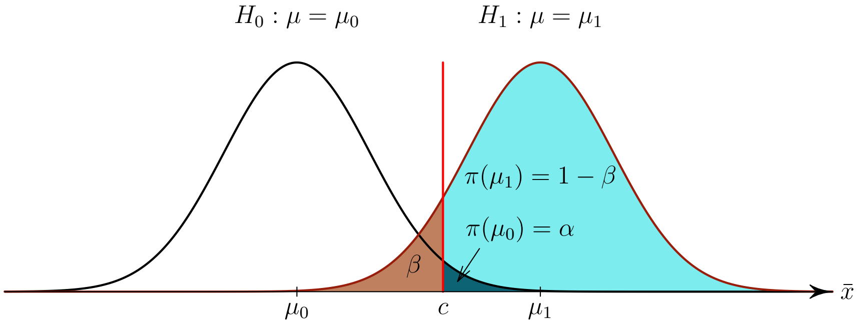

Figure 11.1: The relation ship of the power function to the probability of a Type II error.

Definition 11.1.4: p-value

The **observed size** or **p-value** of the test is defined as the smallest size α at which H_0 can be rejected based on the observed value of the test statistic:

$$\text{p-value} = \inf\{\alpha : (x_1, \ldots, x_n) \in \mathcal{R}_\alpha\}. \tag{11.1.10}$$

11.2 Common Tests

11.2.1 Pivotal Test

Usually, the rejection region $\mathcal{R}$ is of the form

$$\mathcal{R} = \{(x_1, \ldots, x_n) : q_\theta \in (-\infty, k_0] \cup [k_1, \infty)\} \tag{11.2.1}$$

where $Q_\theta = \mathscr{Q}(X_1, \ldots, X_n; \theta)$ is a **test statistic**, $q_\theta = \mathscr{Q}(x_1, \ldots, x_n; \theta)$ is the observed value of Q_θ, and k_1, k_2 are **critical values**. Under H_0, Q_{θ_0} is a **pivotal quantity**, i.e. $Q_{\theta_0} \sim f_{Q_{\theta_0}}(q)$.

Two-Sided Test

A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

is called **two-sided test**. Since p-value is the probability (assuming the null hypothesis is true) of the test statistic being as extreme or more extreme than was observed, therefore

$$\text{p-value} = \int f_{Q_{\theta_0}}(u) \cdot 1\{f_{Q_{\theta_0}}(u) \le f_{Q_{\theta_0}}(q_\theta)\} \, \mathrm{d}u.$$

For a given significance level $\alpha = \alpha_1 + \alpha_2$,

$$\mathbb{P}(Q_{\theta_0} \le k_0) = \alpha_1 \;\Rightarrow\; k_0, \quad \mathbb{P}(Q_{\theta_0} \ge k_1) = \alpha_2 \;\Rightarrow\; k_1.$$

The null hypothesis H_0 is rejected if

$$q_{\theta_0} \in (-\infty, k_0] \cup [k_1, \infty) \quad \text{or} \quad \text{p-value} \le \alpha.$$

The power function of the test is

$$\pi(\theta) = \mathbb{P}(Q_\theta \leq k_0) + \mathbb{P}(Q_\theta \geq k_1).$$

If H_1 is true, $\theta = \theta_1 \neq \theta_0$, then the probability of Type II error is given by

$$\beta(\theta_1) = 1 - \pi(\theta_1) = \mathbb{P}(k_0 < Q_{\theta_1} < k_1).$$

The $100(1-\alpha)\%$ confidence interval for θ is (θ_L, θ_U), where

$$\mathbb{P}(q_{\theta_U} < Q_{\theta_0} < q_{\theta_L}) = 1 - \alpha.$$

Upper One-Sided Test

A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

is called **upper one-sided test**. In this case, $k_0 = -\infty$. The p-value is given by

$$\text{p-value} = \mathbb{P}(Q_{\theta_0} \geq q_{\theta_0}).$$

For a given significance level α,

$$\mathbb{P}(Q_{\theta_0} \geq q_{\theta_0}) = \alpha \quad \Rightarrow \quad k_1.$$

The null hypothesis H_0 is rejected if

$$q_{\theta_0} \geq k_1 \quad \text{or} \quad \text{p-value} \leq \alpha.$$

The power function of the test is

$$\pi(\theta) = \mathbb{P}(Q_\theta \geq k_1).$$

If H_1 is true, $\theta = \theta_1 > \theta_0$, then the probability of Type II error is given by

$$\beta(\theta_1) = 1 - \pi(\theta_1) = \mathbb{P}(Q_{\theta_1} < k_1).$$

The $100(1-\alpha)\%$ confidence interval for θ is (θ_L, ∞), where

$$\mathbb{P}(Q_{\theta_0} < q_{\theta_L}) = 1 - \alpha.$$

Lower One-Sided Test

A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called **lower one-sided test**. In this case, $k_1 = \infty$. The p-value is given by

$$\text{p-value} = \mathbb{P}(Q_{\theta_0} \leq q_{\theta_0}).$$

For a given significance level α,

$$\mathbb{P}(Q_{\theta_0} \leq k_0) = \alpha \quad \Rightarrow \quad k_0.$$

The null hypothesis H_0 is rejected if

$$q_{\theta_0} \leq k_0 \quad \text{or} \quad \text{p-value} \leq \alpha.$$

The power function of the test is

$$\pi(\theta) = \mathbb{P}(Q_\theta \leq k_0).$$

If H_1 is true, $\theta < \theta_0$, then the probability of Type II error is given by

$$\beta(\theta_1) = 1 - \pi(\theta_1) = \mathbb{P}(k_0 < Q_{\theta_1}).$$

The $100(1-\alpha)\%$ confidence interval for θ is $(-\infty, \theta_U)$, where

$$\mathbb{P}(q_{\theta_U} < Q_{\theta_0}) = 1 - \alpha.$$

11.2.2 Conditional Test

Usually, the rejection region $\mathcal{R}$ is of the form

$$\mathcal{R} = \{(x_1, \ldots, x_n) : s \in (-\infty, k_0] \cup [k_1, \infty)\} \tag{11.2.2}$$

where $S = \mathscr{S}(X_1, \ldots, X_n) \sim f_S(s|\theta)$ is a **test statistic**, $s = \mathscr{S}(x_1, \ldots, x_n)$ is the observed value of S, and k_1, k_2 are **critical values**.

Two-Sided Test

A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

is called **two-sided test**. Since p-value is the probability (assuming the null hypothesis is true) of the test statistic being as extreme or more extreme than was observed, therefore

$$\text{p-value} = \int f_S(u|\theta = \theta_0) \cdot 1\{f_S(u|\theta = \theta_0) \leq f_S(s|\theta = \theta_0)\} \, \mathrm{d}u.$$

For a given significance level $\alpha = \alpha_1 + \alpha_2$,

$$\mathbb{P}(S \leq k_0|\theta = \theta_0) = \alpha_1 \Rightarrow k_0, \quad \mathbb{P}(S \geq k_1|\theta = \theta_0) = \alpha_2 \Rightarrow k_1.$$

The null hypothesis H_0 is rejected if

$$s \in (-\infty, k_0] \cup [k_1, \infty) \quad \text{or} \quad \text{p-value} \leq \alpha.$$

The power function of the test is

$$\pi(\theta) = \mathbb{P}(S \leq k_0|\theta) + \mathbb{P}(S \geq k_1|\theta).$$

If H_1 is true, $\theta = \theta_1 \neq \theta_0$, then the probability of Type II error is given by

$$\beta(\theta_1) = 1 - \pi(\theta_1) = \mathbb{P}(k_0 < S < k_1|\theta = \theta_1).$$

The $100(1-\alpha)\%$ confidence interval for θ is (θ_L, θ_U), where

$$\mathbb{P}(S > s|\theta = \theta_U) = 1 - \alpha_1, \quad \mathbb{P}(S < s|\theta = \theta_L) = 1 - \alpha_2.$$

Upper One-Sided Test

A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

is called **upper one-sided test**. In this case, $k_0 = -\infty$. The p-value is given by

$$\text{p-value} = \mathbb{P}(S \geq s | \theta = \theta_0).$$

For a given significance level α,

$$\mathbb{P}(S \geq k_1 | \theta = \theta_0) = \alpha \quad \Rightarrow \quad k_1.$$

The null hypothesis H_0 is rejected if

$$s \geq k_1 \quad \text{or} \quad \text{p-value} \leq \alpha.$$

The power function of the test is

$$\pi(\theta) = \mathbb{P}(S \geq k_1 | \theta).$$

If H_1 is true, $\theta = \theta_1 > \theta_0$, then the probability of Type II error is given by

$$\beta(\theta_1) = 1 - \pi(\theta_1) = \mathbb{P}(S < k_1 | \theta = \theta_1).$$

The $100(1-\alpha)\%$ confidence interval for θ is (θ_L, ∞), where

$$\mathbb{P}(S < s | \theta = \theta_L) = 1 - \alpha.$$

Lower One-Sided Test

A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called **lower one-sided test**. In this case, $k_1 = \infty$. The p-value is given by

$$\text{p-value} = \mathbb{P}(S \leq s | \theta = \theta_0).$$

For a given significance level α,

$$\mathbb{P}(S \leq k_0 | \theta = \theta_0) = \alpha \quad \Rightarrow \quad k_0.$$

The null hypothesis H_0 is rejected if

$$s \leq k_0 \quad \text{or} \quad \text{p-value} \leq \alpha.$$

The power function of the test is

$$\pi(\theta) = \mathbb{P}(S \leq k_0 \theta).$$

If H_1 is true, $\theta = \theta_1 < \theta_0$, then the probability of Type II error is given by

$$\beta(\theta_1) = 1 - \pi(\theta_1) = \mathbb{P}(S > k_0 | \theta = \theta_1).$$

The $100(1-\alpha)\%$ confidence interval for θ is $(-\infty, \theta_U)$, where

$$\mathbb{P}(S > s | \theta = \theta_U) = 1 - \alpha.$$

11.3 Most Powerful Tests

Definition 11.3.1: Most Powerful Tests

A test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ based on a critical region $\mathcal{R}^*$ is said to be a **most powerful test** of size α if

1. $\pi_{\mathcal{R}^*}(\theta_0) = \alpha$, and
2. $\pi_{\mathcal{R}^*}(\theta_1) \geq \pi_{\mathcal{R}}(\theta_1)$ for any other critical region $\mathcal{R}$ of size α (that is $\pi_{\mathcal{R}}(\theta_0) = \alpha$).

Theorem 11.3.1: Neyman-Pearson

Suppose that $X_1, \ldots, X_n$ have joint PDF $f(x_1, \ldots, x_n; \theta)$. Let

$$\lambda(x_1, \ldots, x_n; \theta_0, \theta_1) = \frac{f(x_1, \ldots, x_n; \theta_0)}{f(x_1, \ldots, x_n; \theta_1)}, \tag{11.3.1}$$

and let $\mathcal{R}^*$ be the set

$$\mathcal{R}^* = \{(x_1, \ldots, x_n) : \lambda(x_1, \ldots, x_n; \theta_0, \theta_1) \leq k\}, \tag{11.3.2}$$

where k is a constant such that

$$\mathbb{P}[(X_1, \ldots, X_n) \in \mathcal{R}^*; \theta_0] = \alpha. \tag{11.3.3}$$

Then $\mathcal{R}^*$ is the most powerful critical region of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.

11.4 Uniformly Most Powerful Tests

Definition 11.4.1: Uniformly Most Powerful (UMP) Tests

Let $X_1, \ldots, X_n$ have joint PDF $f(x_1, \ldots, x_n; \theta)$ for $\theta \in \Omega$, and consider hypotheses of the form $H_0 : \theta \in \Omega_0$ versus $H_1 : \theta \in \Omega - \Omega_0$, where Ω_0 is a subset of Ω. A critical region $\mathcal{R}^*$, and the associated test, are said to be **uniformly most powerful** (UMP) if

1. $\max\limits_{\theta \in \Omega_0} \pi_{\mathcal{R}^*}(\theta) = \alpha$, and
2. $\pi_{\mathcal{R}^*}(\theta) \geq \pi_{\mathcal{R}}(\theta)$ for all $\theta \in \Omega - \Omega_0$ and all critical region $\mathcal{R}$ of size α.

Definition 11.4.2: Monotone Likelihood Ratio (MLR)

A joint PDF $f(\boldsymbol{x}; \theta)$ is said to have a **monotone likelihood ratio** (MLR) in the statistic $T = \mathscr{T}(\boldsymbol{X})$ if for any two values of the parameter, $\theta_1 < \theta_2$, the ratio $f(\boldsymbol{x}; \theta_2)/f(\boldsymbol{x}; \theta_1)$ depends on $\boldsymbol{x}$ only through the function $\mathscr{T}(\boldsymbol{x})$, and this ratio is a non-decreasing function of $\mathscr{T}(\boldsymbol{x})$.

Theorem 11.4.1

If a joint PDF $f(\boldsymbol{x};\theta)$ has a MLR in the statistic $T = \mathscr{T}(\boldsymbol{X})$, then a UMP test of size α for

- $H_0 : \theta \le \theta_0$ v.s. $H_1 : \theta > \theta_0$ is to reject H_0 if $\mathscr{T}(\boldsymbol{x}) \ge k$, where $\mathbb{P}(T \ge k;\theta) = \alpha$;
- $H_0 : \theta \ge \theta_0$ v.s. $H_1 : \theta < \theta_0$ is to reject H_0 if $\mathscr{T}(\boldsymbol{x}) \le k$, where $\mathbb{P}(T \le k;\theta) = \alpha$.

Theorem 11.4.2

Suppose that $X_1, \ldots, X_n$ have joint PDF of the form

$$f(\boldsymbol{x};\theta) = c(\theta)h(\boldsymbol{x})\exp[q(\theta)\mathscr{T}(\boldsymbol{x})], \tag{11.4.1}$$

where $q(\theta)$ is an increasing function of θ, then a UMP test of size α for

- $H_0 : \theta \le \theta_0$ v.s. $H_1 : \theta > \theta_0$ is to reject H_0 if $\mathscr{T}(\boldsymbol{x}) \ge k$, where $\mathbb{P}(T \ge k;\theta) = \alpha$;
- $H_0 : \theta \ge \theta_0$ v.s. $H_1 : \theta < \theta_0$ is to reject H_0 if $\mathscr{T}(\boldsymbol{x}) \le k$, where $\mathbb{P}(T \le k;\theta) = \alpha$.

Proof If $\theta_1 < \theta_2$, then $q(\theta_1) < q(\theta_2)$, so that

$$\frac{f(\boldsymbol{x};\theta_2)}{f(\boldsymbol{x};\theta_1)} = \frac{c(\theta_2)}{c(\theta_1)}\exp\{[q(\theta_2) - q(\theta_1)]\mathscr{T}(\boldsymbol{x})\}$$

which is an increasing function of $\mathscr{T}(\boldsymbol{x})$ because $q(\theta_2) - q(\theta_1) > 0$. The theorem follows by the MLR property. □

An obvious application of the theorem occurs when $X_1, \ldots, X_n$ is a random sample from a member of the regular exponential class, say $f(x;\theta) = c(\theta)h(x)\exp[q(\theta)u(x)]$ with $\mathscr{T}(\boldsymbol{x}) = \sum u(x_i)$ and $q(\theta)$ an increasing function of θ.

Unbiased Tests

Definition 11.4.3: Unbiased Test

A test of $H_0 : \theta \in \Omega_0$ versus $H_1 : \theta \in \Omega - \Omega_0$ is **unbiased** if

$$\sup_{\theta \in \Omega - \Omega_0} \pi(\theta) \ge \sup_{\theta \in \Omega_0} \pi(\theta). \tag{11.4.2}$$

In other words, the probability of rejecting H_0 when it is false is at least as large as the probability of rejecting H_0 when it is true.

11.5 Generalized Likelihood Ratio Tests

Definition 11.5.1: Generalized Likelihood Ratio (GLR)

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ have joint PDF $f(\boldsymbol{x};\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Omega}$, and consider the hypothesis $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Omega}_0$ versus $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Omega} - \boldsymbol{\Omega}_0$. The **generalized likelihood ratio**

(GLR) is defined by

$$\lambda(\boldsymbol{x}) = \frac{\sup\limits_{\boldsymbol{\theta}\in\boldsymbol{\Omega}_0} f(\boldsymbol{x};\, \boldsymbol{\theta})}{\sup\limits_{\boldsymbol{\theta}\in\boldsymbol{\Omega}} f(\boldsymbol{x};\, \boldsymbol{\theta})} = \frac{f(\boldsymbol{x};\, \hat{\boldsymbol{\theta}}_0)}{f(\boldsymbol{x};\, \hat{\boldsymbol{\theta}})}, \tag{11.5.1}$$

where $\hat{\boldsymbol{\theta}}$ denotes the usual MLE of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}_0$ denotes the MLE under the restriction that H_0 is true.

Theorem 11.5.1: Wilks's Theorem

Suppose that $\boldsymbol{\Omega}$ is an open set with dimension k, and the dimension of $\boldsymbol{\Omega}_0$ is r $(r < k)$. Under regularity conditions and assuming H_0 is true, then

$$-2\ln\lambda(\boldsymbol{x}) \xrightarrow{d} \chi^2(k-r) \tag{11.5.2}$$

as the sample size tends to infinity.

Note if $\boldsymbol{\Omega}_0$ is a point, then the dimension of $\boldsymbol{\Omega}_0$ is 0.

11.6 Conditional Tests

Theorem 11.6.1

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ have joint PDF of the form

$$f(\boldsymbol{x};\, \theta, \boldsymbol{\kappa}) = c(\theta, \boldsymbol{\kappa})h(\boldsymbol{x})\exp\left[\theta\mathscr{T}(\boldsymbol{x}) + \sum_{i=1}^{m}\kappa_i\mathscr{S}_i(\boldsymbol{x})\right]. \tag{11.6.1}$$

If $S_i = \mathscr{S}_i(\boldsymbol{X})$ for $i = 1, \ldots, m$ and $T = \mathscr{T}(\boldsymbol{X})$, then $S_1, \ldots, S_m$ are jointly sufficient for $\kappa_1, \ldots, \kappa_m$ for each fixed θ, and the conditional PDF $f_{T|\boldsymbol{s}}(t;\, \theta)$ does not depend on $\boldsymbol{\kappa}$. Furthermore,

- A size α test of $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ is to reject H_0 if $\mathscr{T}(\boldsymbol{x}) \ge k(\boldsymbol{s})$ where $\mathbb{P}[T \ge k(\boldsymbol{s})|\boldsymbol{s}] = \alpha$ when $\theta = \theta_0$.
- A size α test of $H_0 : \theta \ge \theta_0$ versus $H_1 : \theta < \theta_0$ is to reject H_0 if $\mathscr{T}(\boldsymbol{x}) \le k(\boldsymbol{s})$ where $\mathbb{P}[T \le k(\boldsymbol{s})|\boldsymbol{s}] = \alpha$ when $\theta = \theta_0$.

11.7 Sequential Tests

A **sequential probability raitio test** (SPRT) is defined in terms of a sequence of such ratios. Specifically, we define

$$\lambda_m = \lambda_m(x_1, \ldots, x_m) = \frac{f(x_1;\, \theta_0)\cdots f(x_m;\, \theta_0)}{f(x_1;\, \theta_1)\cdots f(x_m;\, \theta_1)} \tag{11.7.1}$$

for $m = 1, 2, \ldots$, and adopt the following procedure: Let $k_0 < k_1$ be arbitary positive numbers, and compute λ_1 based on the first observation x_1.

1. If $\lambda_1 \leq k_0$, then reject H_0; if $\lambda_1 \geq k_1$, then accept H_0; and if $k_0 < \lambda_1 < k_1$, then take a second observation x_2 and compute λ_2.
2. If $\lambda_2 \leq k_0$, then reject H_0; if $\lambda_2 \geq k_1$, then accept H_0; and if $k_0 < \lambda_2 < k_1$, then take a third observation x_3 and compute λ_3, and so on.

The idea is to continue taking x_i's as long as the ratio λ_m remains between k_0 and k_1, and to stop as soon as either $\lambda_m \leq k_0$ or $\lambda_m \leq k_1$, rejecting H_0 if $\lambda_m \leq k_0$ and accepting H_0 if $\lambda_m \geq k_1$. The critical region, say $\mathcal{R}$, of the resulting sequential test is the union of the following disjoint sets:

$$\mathcal{R}_n = \{(x_1, \ldots, x_n) : k_0 < \lambda_j < k_1, j = 1, \ldots, n-1, \lambda_n \leq k_0\} \tag{11.7.2}$$

for $n = 1, 2, \ldots$. In other words, if for some n, a point $(x_1, \ldots, x_n)$ is in $\mathcal{R}_n$, then H_0 is rejected for a sample of size n. On the other hand, H_0 is accepted if such a point is in an acceptance region, say $\mathcal{A}$, which is the union of disjoint sets $\mathcal{A}_n$ of the following form:

$$\mathcal{A}_n = \{(x_1, \ldots, x_n) : k_0 < \lambda_j < k_1, j = 1, \ldots, n-1, \lambda_n \geq k_1\}. \tag{11.7.3}$$

In the case of the Neyman-Pearson test for fixed sample size n, the constant k was determined so that the size of the test would be some prescribed α. Now it is necessary to find constants k_0 and k_1 so that the SPRT will have prescribed values α and β for the respective probabilities of Type I and Type II error,

$$\alpha = \mathbb{P}(\text{reject } H_0;\, \theta_0) = \sum_{n=1}^{\infty} \int_{\mathcal{R}_n} L_n(\theta_0)\, \mathrm{d}\boldsymbol{x}, \tag{11.7.4}$$

$$\beta = \mathbb{P}(\text{accept } H_0;\, \theta_1) = \sum_{n=1}^{\infty} \int_{\mathcal{A}_n} L_n(\theta_1)\, \mathrm{d}\boldsymbol{x}, \tag{11.7.5}$$

where $L_n(\theta) = f(x_1; \theta) \cdots f(x_n; \theta)$, and $\mathrm{d}\boldsymbol{x} = \mathrm{d}x_1 \cdots \mathrm{d}x_n$. The constants k_0 and k_1 are solutions of the integral equations (11.7.4) and (11.7.5).

Approximate Sequential Tests

Suppose it is required to perform a sequential test with prescribed probabilities of Type I and Type II errors, α and β, respectively. As noted above, the constants k_0 and k_1 can be obtained by solving the integral equations (11.7.4) and (11.7.5), and exact solutions, in general, will be difficult to achieve. Fortunately, it is possible to obtain approximate solutions that are much easier to compute and rather accurate. If α and β are the exact levels desired, then we define constants

$$k_0^* = \frac{\alpha}{1-\beta}, \quad k_1^* = \frac{1-\alpha}{\beta}. \tag{11.7.6}$$

The following discussion suggests using k_0^* and k_1^* as approximations or k_0 and k_1. Using the above stated property that $N < \infty$ with probability 1 and that $\lambda_n(x_1, \ldots, x_n) \leq k_0$ when

$(x_1, \dots, x_n)$ is in $\mathcal{R}_n$, it follows that

$$\begin{aligned}\alpha = \mathbb{P}(\text{reject } H_0;\, \theta_0) &= \sum_{n=1}^{\infty} \int_{\mathcal{R}_n} L_n(\theta_0)\, \mathrm{d}\boldsymbol{x} \\ &\leq \sum_{n=1}^{\infty} \int_{\mathcal{R}_n} k_0 L_n(\theta_1)\, \mathrm{d}\boldsymbol{x} = k_0 \mathbb{P}(\text{reject } H_0;\, \theta_1) = k_0(1-\beta),\end{aligned}$$

and hence $\alpha/(1-\beta) \leq k_0$. Similarly, because $\lambda_n(x_1, \dots, x_n) \geq k_1$ when $(x_1, \dots, x_n)$ is in $\mathcal{A}_n$, it follows that

$$\begin{aligned}1-\alpha = \mathbb{P}(\text{accept } H_0;\, \theta_0) &= \sum_{n=1}^{\infty} \int_{\mathcal{A}_n} L_n(\theta_0)\, \mathrm{d}\boldsymbol{x} \\ &\geq \sum_{n=1}^{\infty} \int_{\mathcal{A}_n} k_1 L_n(\theta_1)\, \mathrm{d}\boldsymbol{x} = k_1 \mathbb{P}(\text{accept } H_0;\, \theta_1) = k_1\beta,\end{aligned}$$

and hence $k_1 \leq (1-\alpha)/\beta$. These results imply the inequality $k_0^* \leq k_0 < k_1 \leq k_1^*$.

A relationship now will be established between the errors for the exact test and those of the approximate test. Denote by α^* and β^* the actual error sizes of the approximate SPRT based on using the constants k_0^* and k_1^*. Also, denote by $\mathcal{R}_n^*$ and $\mathcal{A}_n^*$ the sets that define, respectively, the critical and acceptance regions for the approximate test based on k_0^* and k_1^*. It follows by an argument similar to that given above that

$$\begin{aligned}\alpha^* &= \sum_{n=1}^{\infty} \int_{\mathcal{R}_n^*} L_n(\theta_0)\, \mathrm{d}\boldsymbol{x} \leq k_0^* \sum_{n=1}^{\infty} \int_{\mathcal{R}_n^*} L_n(\theta_1)\, \mathrm{d}\boldsymbol{x} = \frac{\alpha}{1-\beta}(1-\beta^*), \\ 1-\alpha^* &= \sum_{n=1}^{\infty} \int_{\mathcal{A}_n^*} L_n(\theta_0)\, \mathrm{d}\boldsymbol{x} \geq k_1^* \sum_{n=1}^{\infty} \int_{\mathcal{A}_n^*} L_n(\theta_1)\, \mathrm{d}\boldsymbol{x} = \frac{1-\alpha}{\beta}\beta^*.\end{aligned}$$

It follows that $\alpha^*(1-\beta) \leq \alpha(1-\beta^*)$ and $(1-\alpha)\beta^* \leq (1-\alpha^*)\beta$, and their summation is

$$\alpha^* + \beta^* \leq \alpha + \beta. \tag{11.7.7}$$

Thus, the sum of the errors of the approximate test are bounded above by the sum of the error of the exact test.

Expected Sample Size

We now consider a way of assessing the effectiveness of SPRTs in reducing the amount of sampling relative to tests based on fixed sample sizes. Our criterion involves the expected number of observations required to reach a decision.

As before, we denote by N the number of observations required to reach a decision, either reject H_0 or accept H_0. Theoretically, we might attempt to compute its expectation directly from the definition, but as noted previously the distribution of N is quite complicated and thus we will resort to a different approach. Recall that the test is based on observed values of a sequence of random variables $X_1, \dots, X_n$ which are independent and identically distributed with PDF $f(x;\, \theta)$. Theoretically, we could continue taking observations indefinitely, but according to the sequential procedure defined above, we will terminate as soon as $\lambda_n \leq k_0$ or $\lambda_n \geq k_1$ for some n, and we define N as the first such value n.

We now define a new random variable, say

$$Z = \ln \frac{f(X;\theta_0)}{f(X;\theta_1)} \tag{11.7.8}$$

where $X \sim f(x;\theta)$ for either $\theta = \theta_0$ or θ_1. In a similar manner, we can define a whole sequence of such random variables $Z_1, Z_2, \ldots$, based on the sequence $X_1, X_2, \ldots$ and we also can define a sequence of sums,

$$S_m = \sum_{i=1}^{m} Z_i = \ln[\lambda_m(X_1, \ldots, X_m)], \quad m = 1, 2, \ldots. \tag{11.7.9}$$

It follows that N is the subscript of the first sum S_n such that either $S_n \le \ln k_0$ or $S_n \ge \ln k_1$, and we denote the corresponding sum as $S_N = \sum_{i=1}^{N} Z_i$. It is possible to show that

$$\mathbb{E}(S_N) = \mathbb{E}(N)\mathbb{E}(Z) \tag{11.7.10}$$

when $\mathbb{E}(N) < \infty$. This relationship, which is known as Wald's equation, is useful in deriving an approximation to the expected sample size.

- If the sequential test rejects H_0 at step N, then $S_N \le \ln k_0$, and we would except the sum to be close to $\ln k_0$, because it first dropped below this value at the N-th step.
- Similarly, if the test accepts H_0 at step N, then $S_N \ge \ln k_1$, and we would expect the sum to be close to $\ln k_0$ in this case.

These remarks together with Wald's equation suggest the following approximation:

$$\mathbb{E}(N) = \frac{\mathbb{E}(S_N)}{\mathbb{E}(Z)} \simeq \frac{\ln k_0 \mathbb{P}(\text{reject } H_0) + \ln k_1 \mathbb{P}(\text{accept } H_0)}{\mathbb{E}(Z)}. \tag{11.7.11}$$

By using the approximations $k_0 \simeq k_0^* = \alpha/(1-\beta)$ and $k_1 \simeq k_1^* = (1-\alpha)/\beta$, we obtain the following approximation to expected sample size when H_0 is true:

$$\mathbb{E}(N|\theta_0) \simeq \frac{\alpha \ln[\alpha/(1-\beta)] + (1-\alpha)\ln[(1-\alpha)/\beta]}{\mathbb{E}(Z|\theta_0)}. \tag{11.7.12}$$

Similarly, an approximation when H_1 is true is given by

$$\mathbb{E}(N|\theta_1) \simeq \frac{(1-\beta)\ln[\alpha/(1-\beta)] + \beta \ln[(1-\alpha)/\beta]}{\mathbb{E}(Z|\theta_1)}. \tag{11.7.13}$$

Chapter 12

Contingency Tables and Goodness-of-Fit

12.1 One-Sample Binomial Test

Consider a Bernouli-trial type of situation with two posible outcomes, A_1 and A_2, with $\mathbb{P}(A_1) = p_1$ and $\mathbb{P}(A_2) = p_2 = 1 - p_1$. A random sample of n trials is observed, and we let $o_1 = x$ and $o_2 = n - x$ denote the observed number of outcomes of type A_1 and type A_2, respectively. We wish to test $H_0 : p_1 = p_{10}$ versus $H_1 : p_1 \neq p_{10}$. Under H_0, the expected number of outcomes of each type is $e_1 = np_{10}$ and $e_2 = np_{20} = n(1 - p_{10})$. This situation is illustrated in Table 12.1.

Possible Outcomes	A_1	A_2	Total
Probabilities	p_{10}	p_{20}	1
Expected Outcomes	$e_1 = np_{10}$	$e_2 = np_{20}$	n
Observed Outcomes	$o_1 = x$	$o_2 = n - x$	n

Table 12.1: Values of expected and observed outcomes for a binomial experiment.

According Theorem 6.3.1 (Lindeberg-Lévy CLT),

$$\frac{x - np_{10}}{\sqrt{np_{10}(1 - p_{10})}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{12.1.1}$$

as $n \to \infty$. Thus,

$$\begin{aligned}
\chi^2 = \frac{(x - np_{10})^2}{np_{10}(1 - p_{10})} &= \frac{(x - np_{10})^2}{np_{10}} + \frac{(x - np_{10})^2}{n(1 - p_{10})} \\
&= \frac{(x - np_{10})^2}{np_{10}} + \frac{[(n - x) - n(1 - p_{10})]^2}{n(1 - p_{10})} \\
&= \sum_{j=1}^{2} \frac{(o_j - e_j)^2}{e_j} \xrightarrow{d} \chi^2(1). && (12.1.2)
\end{aligned}$$

An approximate size test of H_0 is to reject H_0 if $\chi^2 > \chi^2_{1-\alpha}(1)$.

12.2 r-Sample Binomial Test

Suppose now that $X_i \sim \mathsf{BIN}(n_i, p_i)$ for $i = 1, \ldots, r$, and we wish to test completely specified hypothesis $H_0 : p_i = p_{i0}$, $i = 1, \ldots, r$, where the p_{i0} are known constants. Now let $o_{i1} = x_i$, and $o_{i2} = n_i - x_i$, denote the observed outcomes in the i-th sample, and let $e_{i1} = n_i p_{i0}$ and $e_{i2} = n_i(1 - p_{i0})$ denote the expected outcomes under H_0. This situation is illustrated in Table 12.2.

Sample	Observed (Expected)		
	A_1	A_2	Total
1	o_{11} (e_{11})	o_{12} (e_{12})	n_1
2	o_{21} (e_{21})	o_{22} (e_{22})	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	o_{r1} (e_{r1})	o_{r2} (e_{r2})	n_r

Table 12.2: Table of r-sample binomial observations.

Because a sum of independent chi-square variables is chi-square distributed, we have approximately

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2(r). \tag{12.2.1}$$

An approximate size test of H_0 is to reject H_0 if $\chi^2 > \chi^2_{1-\alpha}(r)$.

Test of Common p

Perhaps a more common problem is to test whether the p_i are all equal, $H_0 : p_1 = p_2 = \cdots = p_r = p$, where the common value p is not specified. We still have the same $r \times 2$ table of observed outcomes, but the value p must be estimated to estimate the expected numbers under H_0. Under H_0 the MLE of p is the pooled estimate

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{r} x_i = \frac{1}{N}\sum_{i=1}^{r} o_{i1},$$

and $\hat{e}_{i1} = n_i\hat{p}$, $\hat{e}_{i2} = n_i(1 - \hat{p})$, where $N = \sum_{i=1}^{r} n_i$. This situation is illustrated in Table 12.3.

The test statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{2} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2(r-1). \tag{12.2.2}$$

An approximate size test of H_0 is to reject H_0 if $\chi^2 > \chi^2_{1-\alpha}(r-1)$.

Generally, one degree of freedom is lost for each unknown parameter estimated.

12.3 One-Sample Multinomial Test

Suppose now that there are c possible types of outcomes, $A_1, \ldots, A_c$, and in a sample size n let $o_1, \ldots, o_c$ denote the number of observed outcomes of each type. We assume probabilities

Sample	Observed (Expected)		
	A_1	A_2	Total
1	o_{11} $(\hat{e}_{11})$	o_{12} $(\hat{e}_{12})$	n_1
2	o_{21} $(\hat{e}_{21})$	o_{22} $(\hat{e}_{22})$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	o_{r1} $(\hat{e}_{r1})$	o_{r2} $(\hat{e}_{r2})$	n_r
Total	$N\hat{p}$	$N(1-\hat{p})$	N

Table 12.3: Table of r-sample binomial observations (o_{ij}) and estimated expectations ($\hat{e}_{i1} = n_i\hat{p}, \hat{e}_{i2} = n_i(1-\hat{p})$).

$\mathbb{P}(A_j) = p_j$, $j = 1, \ldots, c$, where $\sum_{j=1}^{c} p_j = 1$, and we wish to test the completely specified hypothesis $H_0 : p_j = p_{j0}$, $j = 1, \ldots, c$. Under H_0 the expected values for each type are given by $e_j = np_{j0}$. This situation is illustrated in Table 12.4.

Possible Outcomes	A_1	A_2	$\cdots$	A_c	Total
Probabilities	p_{10}	p_{20}	$\cdots$	p_{c0}	1
Expected Outcomes	$e_1 = np_{10}$	$e_2 = np_{20}$	$\cdots$	$e_c = np_{c0}$	n
Observed Outcomes	o_1	o_2	$\cdots$	o_c	n

Table 12.4: Values of expected and observed outcomes for a multinomial experiment.

The chi-square statistic again provides an appealing and convenient test statistic, where approximately

$$\chi^2 = \sum_{j=1}^{c} \frac{(o_j - e_j)^2}{e_j} \sim \chi^2(c-1). \tag{12.3.1}$$

The degree of freedom is $c-1$ because $c-1$ observed values determine the remaining observed value. An approximate size test of H_0 is to reject H_0 if $\chi^2 > \chi^2_{1-\alpha}(c-1)$.

12.4 r-Sample Multinomial Test

Let $A_1, \ldots, A_c$ denote c possible types of outcomes, and let the probability that an outcome of type A_j wil occur for the i-th population (or i-th sample) be denoted by $p_{j|i}$. Note that $\sum_{j=1}^{c} p_{j|i} = 1$ for each $i = 1, \ldots, r$. Also let o_{ij} denote the observed number of outcomes of type A_j in sample i. For a completely specified $H_0 : p_{j|i} = p_{j|i}^{(0)}$, then $e_{ij} = n_i p_{j|i}^{(0)}$ under H_0. This situation is illustrated in Table 12.5.

Approximately for each i,

$$\chi_i^2 = \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2(c-1). \tag{12.4.1}$$

Sample	Observed (Expected)				
	A_1	A_2	$\cdots$	A_c	Total
1	o_{11} (e_{11})	o_{12} (e_{12})	$\cdots$	o_{1c} (e_{12})	n_1
2	o_{21} (e_{21})	o_{22} (e_{22})	$\cdots$	o_{2c} (e_{22})	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	o_{r1} (e_{r1})	o_{r2} (e_{r2})	$\cdots$	o_{rc} (e_{r2})	n_r

Table 12.5: Table of r-sample multinomial observations (o_{ij}) and expectations $(e_{ij} = n_i p_{j|i}^{(0)})$.

And

$$\chi^2 = \sum_{i=1}^{r} \chi_i^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2(r(c-1)). \tag{12.4.2}$$

Test of Common p_j

The more common problem is to test whether the r multinomial populations are the same without specifying the values of the $p_{j|i}$. Thus we consider

$$H_0 : \ p_{j|1} = \cdots = p_{j|r} = p_j \text{ for } j = 1, \ldots, c$$

We must estimate $c-1$ parameters $p_1, \ldots, p_{c-1}$ which also will determine the estimate of p_c because $\sum_{j=1}^{c} p_j = 1$. Under H_0 the MLE of p_j will be the pooled estimate from the pooled sample of $N = \sum_{i=1}^{r} n_i$ items, which gives

$$\hat{p}_j = \frac{1}{N} \sum_{i=1}^{r} o_{ij} = \frac{c_j}{N}, \tag{12.4.3}$$

where c_j is the j-th column total, and

$$\hat{e}_{ij} = n_i \hat{p}_j = n_i c_j / N. \tag{12.4.4}$$

This situation is illustrated in Table 12.6.

The number of degrees of freedom in this case is $r(c-1) - (c-1) = (r-1)(c-1)$, and approximately

$$\chi^2 = \sum_{i=1}^{r} \chi_i^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2((r-1)(c-1)). \tag{12.4.5}$$

12.5 Test for Independence

Suppose that one factor with c categories is associated with columns and a second factor with r categories is associated with rows in an $r \times c$ contingency table. Let p_{ij} denote the probability that a sampled item is classified in the i-th row category and the j-th column category. Let $p_{i+} = \sum_{j=1}^{c} p_{ij}$ denote the marginal probability that an individual is classified

Sample	Observed (Expected)				
	A_1	A_2	$\cdots$	A_c	Total
1	o_{11} $(\hat{e}_{11})$	o_{12} $(\hat{e}_{12})$	$\cdots$	o_{1c} $(\hat{e}_{12})$	n_1
2	o_{21} $(\hat{e}_{21})$	o_{22} $(\hat{e}_{22})$	$\cdots$	o_{2c} $(\hat{e}_{22})$	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	o_{r1} $(\hat{e}_{r1})$	o_{r2} $(\hat{e}_{r2})$	$\cdots$	o_{rc} $(\hat{e}_{r2})$	n_r
Total	c_1	c_2	$\cdots$	c_c	N

Table 12.6: Table of r-sample multinomial observations (o_{ij}) and estimated expectations ($\hat{e}_{ij} = n_i c_j / N$).

	A_1	A_2	$\cdots$	A_c	Total
B_1	p_{11}	p_{12}	$\cdots$	p_{1c}	p_{1+}
B_2	p_{21}	p_{22}	$\cdots$	p_{2c}	p_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_r	p_{r1}	p_{r2}	$\cdots$	p_{rc}	p_{r+}
Total	p_{+1}	p_{+2}	$\cdots$	p_{+c}	1

Table 12.7: Contingency table of probabilities (p_{ij}).

in the i-th row, and let $p_{+j} = \sum_{i=1}^{r} p_{ij}$ denote the marginal probability that an individual is classified in the j-th column, as illustrated in Table 12.7.

To test independence, we test $H_0 : p_{ij} = p_{i+}p_{+j}$. Let $o_{i+} = \sum_{j=1}^{c} o_{ij}$ and $o_{+j} = \sum_{i=1}^{r} o_{ij}$ denote the row and column totals as before, although the n_i are not fixed before the sample in this case. Let $N = \sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij}$ denote the total number of outcomes. Then $\hat{p}_{i+} = o_{i+}/N$, $\hat{p}_{+j} = o_{+j}/N$, and under H_0 the expected number of outcomes to fall in the (i, j) cell is estimated to be

$$\hat{e}_{ij} = N\hat{p}_{ij} = Np_{i+}p_{+j} = o_{i+}o_{+j}/N. \tag{12.5.1}$$

This situation is illustrated in Table 12.8.

	Observed (Expected)				
	A_1	A_2	$\cdots$	A_c	Σ
B_1	o_{11} $(\hat{e}_{11})$	o_{12} $(\hat{e}_{12})$	$\cdots$	o_{1c} $(\hat{e}_{12})$	o_{1+}
B_2	o_{21} $(\hat{e}_{21})$	o_{22} $(\hat{e}_{22})$	$\cdots$	o_{2c} $(\hat{e}_{22})$	o_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_r	o_{r1} $(\hat{e}_{r1})$	o_{r2} $(\hat{e}_{r2})$	$\cdots$	o_{rc} $(\hat{e}_{r2})$	o_{r+}
Σ	o_{+1}	o_{+2}	$\cdots$	o_{+c}	N

Table 12.8: Contingency table of observations (o_{ij}) and estimated expectations ($\hat{e}_{ij}$).

Approximately,

$$\chi^2 = \sum_{i=1}^{r} \chi_i^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2((r-1)(c-1)). \tag{12.5.2}$$

12.6 Chi-Squared Goodness-of-Fit Test

Suppose we wish to test $H_0 : X \sim f(x)$. Simply divide the sample space into c cells, say $A_1, \dots, A_c$ and let $p_{j0} = \mathbb{P}(X \in A_j)$ where $X \sim f(x)$. Then for a random sample of size n, let o_j denote the number of observations that fall into the j-th cell, and under H_0 the expected number in the j-th cell is $e_j = np_{j0}$.

In some cases there may be a natural choice for the cells or the data may be grouped to begin with; otherwise, artificial cells may be chosen. As a general principle, as many cells as possible should be used to increase the number of degrees of freedom, as long as $e_j \geq 5$ or so is maintained to ensure that the chi-squared approximation is fairly accurate.

<table>
<tr><th></th><th>A_1</th><th>A_2</th><th>A_3</th><th>$\cdots$</th><th>A_c</th><th>Total</th></tr>
<tr><td>o_j</td><td>o_1</td><td>o_2</td><td>o_3</td><td>$\cdots$</td><td>o_c</td><td>n</td></tr>
<tr><td>p_{j0}</td><td>p_{10}</td><td>p_{20}</td><td>p_{20}</td><td>$\cdots$</td><td>p_{c0}</td><td>1</td></tr>
<tr><td>$e_j = np_{j0}$</td><td colspan="2">$\underbrace{e_1 \quad e_2}$</td><td>$e_3$</td><td>$\cdots$</td><td>$e_c$</td><td>$n$</td></tr>
<tr><td>$o_j - e_j$</td><td colspan="2">$*$</td><td>$*$</td><td>$\cdots$</td><td>$*$</td><td>0</td></tr>
<tr><td>$(o_j - e_j)^2/e_j$</td><td colspan="2">$*$</td><td>$*$</td><td>$\cdots$</td><td>$*$</td><td>χ^2</td></tr>
</table>

Table 12.9: Observed and expected frequencies for chi-square goodness-of-fit.

Therefore, we combine cells to satisfy $e_j \geq 5$, as illustrated in Table 12.9. Suppose the number of combined cells is c^*. This is now back in the form of the multinomial problem:

$$\chi^2 = \sum_{j=1}^{c^*} \frac{(o_j - e_j)^2}{e_j} \sim \chi^2(c^* - 1). \tag{12.6.1}$$

Unknown Parameter Case

Suppose we wish to test $H_0 : X \sim f(x; \theta_1, \dots, \theta_k)$, where there are k unknown parameters. To compute the χ^2 statistic, the expected numbers under H_0 now must be estimated. If the original data are grouped into cells, then the joint density of the observed values, o_j, is multinomial where the true but unknown $p_{j0} = \mathbb{P}(X \in A_j)$ are functions of $\theta_1, \dots, \theta_k$. If maximum likelihood estimation (MLE) is used to estimate $\theta_1, \dots, \theta_k$ (based on the multinomial distribution of grouped data values o_j), then the limiting distribution of the χ^2 statistic is chi-squared with degrees of freedom $c^* - 1 - k$, where c^* is the number of combined cells and k is the number of parameters estimated. That is, approximately,

$$\chi^2 = \sum_{j=1}^{c^*} \frac{(o_j - \hat{e}_j)^2}{\hat{e}_j} \sim \chi^2_{1-\alpha}(c^* - 1 - k). \tag{12.6.2}$$

where $\hat{e}_j = n\hat{p}_{j0}$, and $\hat{p}_{j0} = \mathbb{P}(X \in A_j;\, \hat{\theta}_1, \ldots, \hat{\theta}_k)$.

12.7 Fisher's Exact Test

$$p_1 = \mathbb{P}(Y = A|X = 1), \tag{12.7.1a}$$

$$p_2 = \mathbb{P}(Y = A|X = 2). \tag{12.7.1b}$$

We want to test the hypothesis:

$$H_0 : p_1 = p_2 \quad \text{versus} \quad H_1 : p_1 > p_2. \tag{12.7.2}$$

Under H_0, we have $O_1 \sim \mathsf{BIN}(n_1, p)$, $O_2 \sim \mathsf{BIN}(n_2, p)$ and $O \sim \mathsf{BIN}(n, p)$. Therefore,

$$\begin{aligned}
\mathbb{P}(O_1 = o_1|O_1 + O_2 = o) &= \frac{\mathbb{P}(O_1 = o_1,\, O_2 = o - o_1)}{\mathbb{P}(O_1 + O_2 = o)} = \frac{\mathbb{P}(O_1 = o_1)\,\mathbb{P}(O_2 = o - o_1)}{\mathbb{P}(O_1 + O_2 = o)} \\
&= \frac{\left[\binom{n_1}{o_1} p^{o_1}(1-p)^{n_1 - o_1}\right]\left[\binom{n_2}{o - o_1} p^{o - o_1}(1-p)^{n_2 - (o - o_1)}\right]}{\binom{n}{o} p^{o}(1-p)^{n-o}} \\
&= \frac{\binom{n_1}{o_1}\binom{n_2}{o - o_1}}{\binom{n}{o}}.
\end{aligned}$$

	$Y = A$	$Y = B$	Total
$X = 1$	o_1	$n_1 - o_1$	n_1
$X = 2$	o_2	$n_2 - o_2$	n_2
Total	o	$n - o$	n

Table 12.10: Observed frequencies for Fisher's exact test.

Chapter 13

Multivariate Statistics

13.1 Vectors of Random Variables

A random matrix is a matrix of random variables

$$\mathbf{Z} = \begin{bmatrix} Z_{11} & \cdots & Z_{1n} \\ \vdots & \ddots & \vdots \\ Z_{m1} & \cdots & Z_{mn} \end{bmatrix}. \tag{13.1.1}$$

Its expectation is given by

$$\mathbb{E}(\mathbf{Z}) = \begin{bmatrix} \mathbb{E}(Z_{11}) & \cdots & \mathbb{E}(Z_{1n}) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(Z_{m1}) & \cdots & \mathbb{E}(Z_{mn}) \end{bmatrix}. \tag{13.1.2}$$

Theorem 13.1.1

If $\mathbf{A} \in \mathbb{R}^{l\times m}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$ and $\mathbf{C} \in \mathbb{R}^{l\times p}$ are matrices, respectively, of constants, then

$$\mathbb{E}(\mathbf{AZB} + \mathbf{C}) = \mathbf{A}\mathbb{E}(\mathbf{Z})\mathbf{B} + \mathbf{C}. \tag{13.1.3}$$

Proof Let $\mathbf{W} = \mathbf{AZB} + \mathbf{C}$, then

$$W_{ij} = \sum_{r=1}^{m}\sum_{s=1}^{n} a_{ir}Z_{rs}b_{sj} + c_{ij},$$

and

$$\mathbb{E}(W_{ij}) = \sum_{r=1}^{m}\sum_{s=1}^{n} a_{ir}\mathbb{E}(Z_{rs})b_{sj} + c_{ij} = (\mathbf{AZB})_{ij} + c_{ij}. \qquad \square$$

Using similar algebra, we can prove the following theorem.

Theorem 13.1.2

If $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{m\times n}$ are matrices of constants, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $n \times 1$ vectors of random variables, then

$$\mathbb{E}(\mathbf{A}\boldsymbol{X} + \mathbf{B}\boldsymbol{Y}) = \mathbf{A}\mathbb{E}(\boldsymbol{X}) + \mathbf{B}\mathbb{E}(\boldsymbol{Y}). \tag{13.1.4}$$

Definition 13.1.1: Covariance

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $m \times 1$ and $n \times 1$ vectors of random variables, then the covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \begin{bmatrix} \operatorname{Cov}(X_1, Y_1) & \cdots & \operatorname{Cov}(X_1, Y_n) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_m, Y_1) & \cdots & \operatorname{Cov}(X_m, Y_n) \end{bmatrix}. \tag{13.1.5}$$

Theorem 13.1.3

If $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\alpha}$ and $\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\beta}$, then

$$\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\alpha})(\boldsymbol{Y} - \boldsymbol{\beta})^\top\right]. \tag{13.1.6}$$

Proof

$$\operatorname{Cov}(X_i, Y_j) = \mathbb{E}[(X_i - \alpha_i)(Y_j - \beta_j)] = \left\{\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\alpha})(\boldsymbol{Y} - \boldsymbol{\beta})^\top\right]\right\}_{ij}. \qquad \square$$

Theorem 13.1.4

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $m \times 1$ and $n \times 1$ vectors of random variables, and $\mathbf{A} \in \mathbb{R}^{l\times m}$ and $\mathbf{B} \in \mathbb{R}^{p\times n}$ are matrices of constants, respectively, then

$$\operatorname{Cov}(\mathbf{A}\boldsymbol{X}, \mathbf{B}\boldsymbol{Y}) = \mathbf{A}\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y})\mathbf{B}^\top. \tag{13.1.7}$$

Proof

$$\begin{aligned}\operatorname{Cov}(\mathbf{A}\boldsymbol{X}, \mathbf{B}\boldsymbol{Y}) &= \mathbb{E}\left\{[\mathbf{A}\boldsymbol{X} - \mathbb{E}(\mathbf{A}\boldsymbol{X})]\,[\mathbf{B}\boldsymbol{Y} - \mathbb{E}(\mathbf{B}\boldsymbol{Y})]^\top\right\} \\ &= \mathbb{E}\left[\mathbf{A}(\boldsymbol{X} - \boldsymbol{\alpha})(\boldsymbol{Y} - \boldsymbol{\beta})\mathbf{B}^\top\right] \\ &= \mathbf{A}\mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\alpha})(\boldsymbol{Y} - \boldsymbol{\beta})\right]\mathbf{B}^\top = \mathbf{A}\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y})\mathbf{B}^\top. \qquad \square\end{aligned}$$

Theorem 13.1.5

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $m \times 1$ and $n \times 1$ vectors of random variables, and $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^n$ are vectors of constants, then

$$\operatorname{Cov}(\boldsymbol{X} - \boldsymbol{a}, \boldsymbol{Y} - \boldsymbol{b}) = \operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y}). \tag{13.1.8}$$

Definition 13.1.2: Variance

If $\boldsymbol{X}$ is a $n \times 1$ vector of random variables, then the variance of $\boldsymbol{X}$ is defined as

$$\operatorname{Var}(\boldsymbol{X}) = \operatorname{Cov}(\boldsymbol{X}, \boldsymbol{X}) = \begin{bmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_n, X_1) & \operatorname{Cov}(X_n, X_2) & \cdots & \operatorname{Var}(X_n) \end{bmatrix}. \tag{13.1.9}$$

Theorem 13.1.6

If $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, then

$$\operatorname{Var}(\boldsymbol{X}) = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right] = \mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}^\top\right) - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{13.1.10}$$

Theorem 13.1.7

$\operatorname{Var}(\boldsymbol{X})$ is symmetric, and positive-semidefinite.

Proof Since $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$, $\operatorname{Var}(\boldsymbol{X})$ is symmetric. For any vector $\boldsymbol{c}$,

$$Q(\boldsymbol{c}) = \boldsymbol{c}^\top \operatorname{Var}(\boldsymbol{X})\boldsymbol{c} = \boldsymbol{c}^\top \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right] \boldsymbol{c} = \operatorname{Var}[\boldsymbol{c}^\top(\boldsymbol{X} - \boldsymbol{\mu})] \geq 0.$$

□

Theorem 13.1.8

If $\boldsymbol{X}$ is a vector of random variables such that no element of $\boldsymbol{X}$ is a linear combination of the remaining elements [i.e., there do not exist $\boldsymbol{a}$ ($\neq \boldsymbol{0}$) and $\boldsymbol{b}$ such that $\boldsymbol{a}^\top \boldsymbol{X} = \boldsymbol{b}$ for all values of $\boldsymbol{X} = \boldsymbol{x}$], then $\operatorname{Var}(\boldsymbol{X})$ is a positive-definite matrix.

Proof For any vector $\boldsymbol{c}$, we have

$$\begin{aligned} 0 &\leq \operatorname{Var}\left(\boldsymbol{c}^\top \boldsymbol{X}\right) \\ &= \boldsymbol{c}^\top \operatorname{Var}(\boldsymbol{X})\boldsymbol{c} && \text{(by eq.(13.1.7))} \end{aligned}$$

Now equality holds iff $\boldsymbol{c}^\top \boldsymbol{X}$ is a constant, that is, iff $\boldsymbol{c}^\top \boldsymbol{X} = \boldsymbol{d}$ ($\boldsymbol{c} \neq \boldsymbol{0}$) or $\boldsymbol{c} = \boldsymbol{0}$. Because the former possibility is ruled out, $\boldsymbol{c} = \boldsymbol{0}$. Thus, $\operatorname{Var}(\boldsymbol{X})$ is positive-definite. □

Definition 13.1.3: Moment Generating Function

If $\boldsymbol{X}$ and $\boldsymbol{t}$ are $n \times 1$ vectors of random variables and constants, respectively, then the **moment generating function** (MGF) of $\boldsymbol{X}$ is defined to be

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}\left[\exp\left(\boldsymbol{t}^\top \boldsymbol{X}\right)\right]. \tag{13.1.11}$$

13.2 Quadratic Form

Theorem 13.2.1: Expectation of a Quadratic Form

Let $\boldsymbol{X}$ be an $n \times 1$ vector of random variables, and let $\mathbf{A}$ be an $n \times n$ symmetric matrix. If $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\operatorname{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, then

$$\mathbb{E}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right) = \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}. \tag{13.2.1}$$

Proof **Method I**:

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right) &= \mathbb{E}\left(\sum_{i=1}^n \sum_{j=1}^n a_{ij} X_i X_j\right) \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mathbb{E}(X_i X_j) && \text{(by linearity of expectation)} \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ij} (\sigma_{ij} + \mu_i \mu_j) && \text{(apply covariance formula)} \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \sigma_{ji} + \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mu_i \mu_j && \text{(since } \boldsymbol{\Sigma} \text{ is a symmetric matrix)} \\
&= \sum_{i=1}^n [\mathbf{A}\boldsymbol{\Sigma}]_{ii} + \boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu} \\
&= \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}.
\end{aligned}$$

Method II: Since the quadratic form is a scalar quantity,

$$\mathbb{E}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right) = \operatorname{tr}\left[\mathbb{E}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)\right].$$

Since the trace operator is a linear combination of the components of the matrix, it therefore follows from the linearity of the expectation operator that

$$\operatorname{tr}\left[\mathbb{E}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)\right] = \mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)\right].$$

Next, by the cyclic property of the trace operator,

$$\mathbb{E}\left[\operatorname{tr}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)\right] = \mathbb{E}\left[\operatorname{tr}\left(\mathbf{A}\boldsymbol{X}\boldsymbol{X}^\top\right)\right].$$

Another application of linearity of expectation tells us that

$$\mathbb{E}\left[\operatorname{tr}\left(\mathbf{A}\boldsymbol{X}\boldsymbol{X}^\top\right)\right] = \operatorname{tr}\left[\mathbb{E}\left(\mathbf{A}\boldsymbol{X}\boldsymbol{X}^\top\right)\right] = \operatorname{tr}\left[\mathbf{A}\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)\right].$$

A standard property of variances then tells us that this is

$$\operatorname{tr}\left[\mathbf{A}\mathbb{E}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)\right] = \operatorname{tr}\left[\mathbf{A}\left(\operatorname{Var}(\boldsymbol{X}) + \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\right] = \operatorname{tr}\left(\mathbf{A}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(\mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}^\top\right).$$

Applying the cyclic property of the trace operator again, we get the result desired. □

Example 13.2.1

Suppose that the elements of $\boldsymbol{X} = (X_1, \dots, X_n)^\top$ have a common mean μ and $\boldsymbol{X}$ has variance matrix $\boldsymbol{\Sigma}$ with $\sigma_{ii} = \sigma^2$ and $\sigma_{ij} = \rho\sigma^2$ $(i \neq j)$. Then, to find the expected value of $Q = \sum_i (X_i - \bar{X})^2$, we express Q in the form $Q = \boldsymbol{X}^\top \mathbf{A} \boldsymbol{X}$, where[a]

$$\mathbf{A} = \mathbf{I}_n - n^{-1}\mathbf{J}_n = \begin{bmatrix} 1 - n^{-1} & -n^{-1} & \cdots & -n^{-1} \\ -n^{-1} & 1 - n^{-1} & \cdots & -n^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ -n^{-1} & -n^{-1} & \cdots & 1 - n^{-1} \end{bmatrix}. \tag{13.2.2}$$

Since

$$\boldsymbol{\Sigma} = \sigma^2 \left[(1-\rho)\mathbf{I}_n + \rho \mathbf{J}_n\right] = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \tag{13.2.3}$$

we have

$$\begin{aligned} \mathbf{A}\boldsymbol{\Sigma} &= \sigma^2 \left[\mathbf{I}_n - n^{-1}\mathbf{J}_n\right] \left[(1-\rho)\mathbf{I}_n + \rho \mathbf{J}_n\right] \\ &= \sigma^2 \left[(1-\rho)\mathbf{I}_n - (1-\rho)n^{-1}\mathbf{J}_n + \rho\mathbf{J}_n - \rho\mathbf{J}_n\right] \\ &= \sigma^2(1-\rho)\mathbf{A}. \end{aligned}$$

Thus,

$$\mathbb{E}(Q) = \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) = \sigma^2(1-\rho)\operatorname{tr}(\mathbf{A}) = \sigma^2(1-\rho)(n-1).$$

[a] $\mathbf{J}_n$ is an $n \times n$ matrix of ones, i.e. $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^\top$. $\mathbf{J}_n^k = n^{k-1}\mathbf{J}_n$ for $k = 1, 2, \dots$.

Theorem 13.2.2: Variance of a Quadratic Form

Let $X_1, \dots, X_n$ be independent random variables with means $\theta_1, \dots, \theta_n$, common variance μ_2, and common third and fourth moments about their means, μ_3 and μ_4, respectively (i.e., $\mu_r = \mathbb{E}[(X_i - \theta_i)^r)]$. If $\mathbf{A}$ is any $n \times n$ symmetric matrix and $\boldsymbol{d} = \operatorname{diag}(\mathbf{A})$, then

$$\operatorname{Var}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right) = \left(\mu_4 - 3\mu_2^2\right)\boldsymbol{d}^\top\boldsymbol{d} + 2\mu_2^2 \operatorname{tr}\left(\mathbf{A}^2\right) + 4\mu_2\boldsymbol{\theta}^\top\mathbf{A}^2\boldsymbol{\theta} + 4\mu_3\boldsymbol{\theta}^\top\mathbf{A}\boldsymbol{d}. \tag{13.2.4}$$

Proof We note that $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\theta}$, $\operatorname{Var}(\boldsymbol{X}) = \mu_2\mathbf{I}_n$, and

$$\operatorname{Var}\left(\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}\right) = \mathbb{E}\left[\left(\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}\right)^2\right] - \left[\mathbb{E}\left(\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}\right)\right]^2.$$

Since $\mathbf{A}$ is symmetric, we have

$$\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X} = (\boldsymbol{X} - \boldsymbol{\theta})^\top\mathbf{A}(\boldsymbol{X} - \boldsymbol{\theta}) + 2\boldsymbol{\theta}^\top\mathbf{A}(\boldsymbol{X} - \boldsymbol{\theta}) + \boldsymbol{\theta}^\top\mathbf{A}\boldsymbol{\theta},$$

so that squaring give

$$\begin{aligned}\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)^2 &= \left[(\boldsymbol{X}-\boldsymbol{\theta})^\top \mathbf{A}(\boldsymbol{X}-\boldsymbol{\theta})\right]^2 + 4\left[\boldsymbol{\theta}^\top \mathbf{A}(\boldsymbol{X}-\boldsymbol{\theta})\right]^2 + \left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\right)^2\\ &\quad + 2\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\left[(\boldsymbol{X}-\boldsymbol{\theta})^\top \mathbf{A}(\boldsymbol{X}-\boldsymbol{\theta}) + 4\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbf{A}(\boldsymbol{X}-\boldsymbol{\theta})\right]\\ &\quad + 4\boldsymbol{\theta}^\top \mathbf{A}(\boldsymbol{X}-\boldsymbol{\theta})(\boldsymbol{X}-\boldsymbol{\theta})^\top \mathbf{A}(\boldsymbol{X}-\boldsymbol{\theta}).\end{aligned}$$

Setting $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\theta}$, we have $\mathbb{E}(\boldsymbol{Y}) = \mathbf{0}$, and

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\right)^2\right] &= \mathbb{E}\left[\left(\boldsymbol{Y}^\top \mathbf{A}\boldsymbol{Y}\right)^2\right] + 4\mathbb{E}\left[\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{Y}\right)^2\right] + \mathbb{E}\left[\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\right)^2\right]\\ &\quad + 2\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta}\mu_2 \operatorname{tr}(\mathbf{A}) + 4\mathbb{E}\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{Y}\boldsymbol{Y}^\top \mathbf{A}\boldsymbol{Y}\right).\end{aligned}$$

As a first step in evaluating the expression above we note that

$$\left(\boldsymbol{Y}^\top \mathbf{A}\boldsymbol{Y}\right)^2 = \sum_i\sum_j\sum_k\sum_l a_{ij}a_{kl}Y_iY_jY_kY_l.$$

Since the Y_i are mutually independent with the same first four moments about the origin, we have

$$\mathbb{E}[Y_iY_jY_kY_l] = \begin{cases}\mu_4, & i=j=k=l,\\ \mu_2, & i=j, k=l;\ \text{or } i=k, j=l;\ \text{or } i=l, j=k,\\ 0, & \text{otherwise.}\end{cases}$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{Y}^\top \mathbf{A}\boldsymbol{Y}\right)^2\right] &= \mu_4\sum_i a_{ii}^2 + \mu_2\sum_i\left(\sum_{k\neq i} a_{ii}a_{kk} + \sum_{j\neq i} a_{ij}^2 + \sum_{j\neq i} a_{ij}a_{ji}\right)\\ &= \left(\mu_4 - 3\mu_2^2\right)\boldsymbol{d}^\top \boldsymbol{d} + \mu_2^2\left[\operatorname{tr}(\mathbf{A})^2 + 2\operatorname{tr}\left(\mathbf{A}^2\right)\right],\end{aligned}$$

since $\mathbf{A}$ is symmetric and $\sum_i\sum_j a_{ij}^2 = \sum_i\sum_j a_{ij}a_{ji} = \operatorname{tr}\left(\mathbf{A}^2\right)$. Let $\boldsymbol{b} = \mathbf{A}\boldsymbol{\theta}$, then

$$\begin{aligned}\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{Y}\right)^2 = \left(\boldsymbol{b}^\top \boldsymbol{Y}\right)^2 &= \sum_i\sum_j b_ib_jY_iY_j,\\ \Rightarrow \mathbb{E}\left[\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{Y}\right)^2\right] &= \mu_2\sum_i b_i^2 = \mu_2\boldsymbol{b}^\top \boldsymbol{b} = \mu_2\boldsymbol{\theta}\mathbf{A}^2\boldsymbol{\theta}.\\ \boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{Y}\boldsymbol{Y}^\top \mathbf{A}\boldsymbol{Y} &= \sum_i\sum_j\sum_k b_ia_{jk}Y_iY_jY_k,\\ \Rightarrow \mathbb{E}\left(\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{Y}\boldsymbol{Y}^\top \mathbf{A}\boldsymbol{Y}\right) &= \mu_3\sum_i b_ia_{ii} = \mu_3\boldsymbol{b}^\top \boldsymbol{d} = \mu_3\sum_i b_ia_{ii} = \mu_3\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{d}.\end{aligned}$$

Finally, collecting all the terms leads to the desired result. □

13.3 Multivariate Normal Distribution

Definition 13.3.1: Multivariate Normal Distribution

A random vector $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ is said to have the **multivariate normal distribution** if it satisfies the following equivalent conditions:

I For every n-vector $\boldsymbol{a}$, the random variable $Y = \boldsymbol{a}^\top \boldsymbol{X}$ has a normal distribution.

II There is an n-vector $\boldsymbol{\mu}$ and a symmetric positive-semidefinite $n \times n$ matrix $\boldsymbol{\Sigma}$, such that the MGF of $\boldsymbol{X}$ is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma} \boldsymbol{t}\right). \tag{13.3.1}$$

III There exists a random l-vector $\boldsymbol{Z}$, whose components are independent standard normal random variables, an n-vector $\boldsymbol{\mu}$, and a $n \times l$ matrix $\mathbf{A}$, such that $\boldsymbol{X} = \mathbf{A}\boldsymbol{Z} + \boldsymbol{\mu}$. Here the covariance matrix $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$.

It is denoted by $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X})$ is the mean vector, and $\boldsymbol{\Sigma} = \operatorname{Var}(\boldsymbol{X})$ is the variance matrix.

Theorem 13.3.1

If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y} = \mathbf{C}\boldsymbol{X} + \boldsymbol{d} \sim \mathcal{N}\left(\mathbf{C}\boldsymbol{\mu} + \boldsymbol{d}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top\right)$.

Proof For every vector $\boldsymbol{a}$,

$$\boldsymbol{a}^\top \boldsymbol{Y} = \boldsymbol{a}^\top \mathbf{C}\boldsymbol{X} + \boldsymbol{a}^\top \boldsymbol{d} = \left(\mathbf{C}^\top \boldsymbol{a}\right)^\top \boldsymbol{X} + \boldsymbol{a}^\top \boldsymbol{d} = \boldsymbol{b}^\top \boldsymbol{X} + c,$$

where $\boldsymbol{b} = \mathbf{C}^\top \boldsymbol{a}$ and $c = \boldsymbol{a}^\top \boldsymbol{d}$. Since $\boldsymbol{b}^\top \boldsymbol{X}$ is normal according to Definition 13.3.1 I (and c is a constant), it follows that $\boldsymbol{a}^\top \boldsymbol{Y}$ is normal. And

$$\begin{aligned}\mathbb{E}(\boldsymbol{Y}) &= \mathbf{C}\mathbb{E}(\boldsymbol{X}) + \boldsymbol{d} = \mathbf{C}\boldsymbol{\mu} + \boldsymbol{d}, \\ \operatorname{Var}(\boldsymbol{Y}) &= \operatorname{Var}(\mathbf{C}\boldsymbol{X} + \boldsymbol{d}) = \operatorname{Var}(\mathbf{C}\boldsymbol{X}) = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top.\end{aligned}$$ □

Corollary 13.3.1

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, $\mathbf{Q}$ is an $n \times n$ orthogonal matrix ($\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_n$), and $\boldsymbol{Y} = \mathbf{Q}\boldsymbol{X}$. Then, $Y_1, \ldots, Y_n$ are mutually independent variables, and $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{Q}\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$.

Lemma 13.3.1

If $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $M_{\boldsymbol{Z}}(\boldsymbol{t}) = \exp\left(\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{t}\right)$.

Proof

$$\begin{aligned}M_{\boldsymbol{Z}}(\boldsymbol{t}) &= \mathbb{E}\left[\exp\left(\boldsymbol{t}^\top \boldsymbol{Z}\right)\right] = \mathbb{E}\left[\exp\left(\sum_{i=1}^n t_i Z_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^n \exp\left(t_i Z_i\right)\right] \\ &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(t_i Z_i\right)\right] = \prod_{i=1}^n \exp\left(\frac{1}{2}t_i^2\right) = \exp\left(\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{t}\right).\end{aligned}$$ □

Theorem 13.3.2

In Definition 13.3.1, I and II are equivalent.

Proof $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow$ II: Because $\boldsymbol{\Sigma}$ is symmetric positive-semidefinite, it has a symmetric positive-semidefinite square root $\boldsymbol{\Sigma}^{1/2}$, which satisfies $\left(\boldsymbol{\Sigma}^{1/2}\right)^2 = \boldsymbol{\Sigma}$. Let $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$, then, by Theorem 13.3.1, we have $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and

$$\begin{aligned}
M_{\boldsymbol{X}}(\boldsymbol{t}) &= \mathbb{E}\left[\exp\left(\boldsymbol{t}^\top \boldsymbol{X}\right)\right] = \mathbb{E}\left[\exp\left(\boldsymbol{t}^\top \boldsymbol{\mu} + \boldsymbol{t}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\right)\right] \\
&= \exp\left(\boldsymbol{t}^\top \boldsymbol{\mu}\right) \cdot \mathbb{E}\left\{\exp\left[\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}\right)^\top \boldsymbol{Z}\right]\right\} \\
&= \exp\left(\boldsymbol{t}^\top \boldsymbol{\mu}\right) \cdot M_{\boldsymbol{Z}}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}\right) \\
&= \exp\left(\boldsymbol{t}^\top \boldsymbol{\mu}\right) \cdot \exp\left[\frac{1}{2}\left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}\right)^\top \left(\boldsymbol{\Sigma}^{1/2}\boldsymbol{t}\right)\right] \qquad \text{(by Lemma 13.3.1)} \\
&= \exp\left(\boldsymbol{t}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{\Sigma}\boldsymbol{t}\right).
\end{aligned}$$

II $\Rightarrow \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: Let $\boldsymbol{a}$ be an arbitrary n-vector, $Y = \boldsymbol{a}^\top \boldsymbol{X}$, then

$$\begin{aligned}
M_Y(u) = \mathbb{E}[\exp(uY)] = \mathbb{E}\left[\exp\left(u\boldsymbol{a}^\top \boldsymbol{X}\right)\right] &= M_{\boldsymbol{X}}(u\boldsymbol{a}) \\
= \exp\left[(u\boldsymbol{a})^\top \boldsymbol{\mu} + \frac{1}{2}(u\boldsymbol{a})^\top \boldsymbol{\Sigma}(u\boldsymbol{a})\right] &= \exp\left(um + \frac{1}{2}u^2\sigma^2\right),
\end{aligned}$$

where $m = \boldsymbol{a}^\top \boldsymbol{\mu}$ and $\sigma^2 = \boldsymbol{a}^\top \boldsymbol{\Sigma}\boldsymbol{a}$, which proves that $Y \sim \mathcal{N}(m, \sigma^2)$ and hence that $\boldsymbol{X}$ is normal in the sense of Definition I. □

Theorem 13.3.3

Let $\boldsymbol{\Sigma}$ be a symmetric positive-definite matrix and $\boldsymbol{\mu}$ an n-vector. Then $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ iff $\boldsymbol{X}$ has the PDF given by

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{13.3.2}$$

Proof "$\Leftarrow$": Let $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})$, so that $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}$. The Jacobian of this transformation is

$$J = \det\left(\frac{\partial x_i}{\partial z_j}\right) = \det\left(\boldsymbol{\Sigma}^{1/2}\right) = [\det(\boldsymbol{\Sigma})]^{1/2},$$

thus $\boldsymbol{z}$ has PDF:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f(\boldsymbol{x}(\boldsymbol{z}))\,|J| = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z_i^2\right).$$

The factorization of the joint PDF implies Z_i are mutually independent normal variables and $Z_i \sim \mathcal{N}(0, 1)$. By Theorem 13.3.1, the result is obtained.

The proof of "$\Rightarrow$" is similar. □

13.4 Statistic Independence

Theorem 13.4.1

Let $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and partition $\boldsymbol{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{13.4.1}$$

Then:

1. $\boldsymbol{X}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.

2. $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \mathbf{O}_{p\times(n-p)}$.

3. $\boldsymbol{X}_1 \mid \boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$.

Proof 1. Writing $\boldsymbol{X}_1 = \mathbf{B}\boldsymbol{X}$, where $\mathbf{B} = [\mathbf{I}_p, \mathbf{O}_{p\times(n-p)}]$. Then $\mathbf{B}\boldsymbol{\mu} = \boldsymbol{\mu}_1$ and $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top = \boldsymbol{\Sigma}_{11}$, so the result follows from Theorem 13.3.1.

2. The MGF of $\boldsymbol{X}$ is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^\top\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}\right) = \exp\left(\sum_{i=1}^{2}\boldsymbol{t}_i^\top\boldsymbol{\mu}_i + \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\boldsymbol{t}_i^\top\boldsymbol{\Sigma}_{ij}\boldsymbol{t}_j\right).$$

If $\boldsymbol{\Sigma}_{12} = \mathbf{O}$, the exponent can be written as a function of just $\boldsymbol{t}_1$ plus a function of just $\boldsymbol{t}_2$, so the MGF factorizes into a term in $\boldsymbol{t}_1$ alone times a term in $\boldsymbol{t}_2$ alone. This implies that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent.

Conversely, if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent, then

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{i=1}^{2} M_{\boldsymbol{X}_i}(\boldsymbol{t}_i) = \prod_{i=1}^{2}\exp\left(\boldsymbol{t}_i^\top\boldsymbol{\mu}_i + \frac{1}{2}\boldsymbol{t}_i^\top\boldsymbol{\Sigma}_{ii}\boldsymbol{t}_i\right),$$

which implies $\boldsymbol{t}_1^\top\boldsymbol{\Sigma}_{12}\boldsymbol{t}_2 = 0$ for all $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$, which in turn implies $\boldsymbol{\Sigma}_{12} = \mathbf{O}$.

3. Lu & Shiou (2002) indicates if $\mathbf{A}$ and $\mathbf{D}$ are symmetric,

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\mathsf{T} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{H}^{-1} & -\mathbf{H}^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{B}^\mathsf{T}\mathbf{H}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{B}^\mathsf{T}\mathbf{H}^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}^{-1}, \quad \mathbf{H} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^\mathsf{T}.$$

By Theorem 13.3.3 and result of 1,

$$f(\boldsymbol{x}_1 \mid \boldsymbol{x}_2) = \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f(\boldsymbol{x}_2)} = \frac{f(\boldsymbol{x})}{f(\boldsymbol{x}_2)} \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{p}_1)^\mathsf{T}\mathbf{Q}^{-1}(\boldsymbol{x}_1 - \boldsymbol{p}_1)\right),$$

where $\boldsymbol{p}_1 = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$, and $\mathbf{Q} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. □

Theorem 13.4.2

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and define $\boldsymbol{U} = \mathbf{A}\boldsymbol{X}$, $\boldsymbol{V} = \mathbf{B}\boldsymbol{X}$. Then $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent iff $\operatorname{Cov}(\boldsymbol{U}, \boldsymbol{V}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top = \mathbf{O}$.

Proof Consider

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{V} \end{pmatrix} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \boldsymbol{X}.$$

Then, by Theorem 13.3.1, the random vector $\boldsymbol{W}$ is multivariate normal with

$$\operatorname{Var}(\boldsymbol{W}) = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix} \operatorname{Var}(\boldsymbol{X}) \begin{bmatrix} \mathbf{A}^\top & \mathbf{B}^\top \end{bmatrix} = \begin{bmatrix} \mathbf{A\Sigma A}^\top & \mathbf{A\Sigma B}^\top \\ \mathbf{B\Sigma A}^\top & \mathbf{B\Sigma B}^\top \end{bmatrix}.$$

Thus, the result follows from Theorem 13.4.1.

13.5 Distribution of Quadratic Forms

Theorem 13.5.1

Let $\mathbf{A}$ be a symmetric matrix. Then $\mathbf{A}$ is idempotent ($\mathbf{A}^2 = \mathbf{A}$) and $\operatorname{rank}(\mathbf{A}) = r$ iff it has r eigenvalues equal to 1 and the $n - r$ eigenvalues equal to 0.

Proof Given $\mathbf{A}^2 = \mathbf{A}$, the $\mathbf{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ implies that

$$\lambda\boldsymbol{x}^\top\boldsymbol{x} = \boldsymbol{x}^\top\mathbf{A}\boldsymbol{x} = \boldsymbol{x}^\top\mathbf{A}^2\boldsymbol{x} = (\mathbf{A}\boldsymbol{x})^\top(\mathbf{A}\boldsymbol{x}) = \lambda^2\boldsymbol{x}^\top\boldsymbol{x},$$

and $\lambda(\lambda - 1) = 0$. Hence the eigenvalues are 0 or 1 and, since $\operatorname{rank}(\mathbf{A}) = r$, $\mathbf{A}$ as r eigenvalues equal to 1 and the $n - r$ eigenvalues equal to 0.

Conversely, if the eigenvalues are 0 or 1, then we can assume without loss of generality that the first r eigenvalues are unity. Hence there exists an orthogonal matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^\top\mathbf{A}\mathbf{Q} = \begin{bmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} = \mathbf{\Lambda}, \quad \text{or } \mathbf{A} = \mathbf{Q\Lambda Q}^\top.$$

Therefore,

$$\mathbf{A}^2 = \mathbf{Q\Lambda Q}^\top\mathbf{Q\Lambda Q}^\top = \mathbf{Q\Lambda}^2\mathbf{Q}^\top = \mathbf{Q\Lambda Q}^\top = \mathbf{A},$$

and $\operatorname{rank}(\mathbf{A}) = r$. □

Theorem 13.5.2

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ and let $\mathbf{A}$ be a symmetric matrix. Then $Y = \sigma^{-2}\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X} \sim \chi^2(r, \delta)$, where $\delta = \sigma^{-2}\boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}$, iff $\mathbf{A}$ is idempotent of rank r.

Proof Since $\mathbf{A}$ is an symmetric matrix, then we can diagonalize it with an orthogonal transformation; that is, there is an orthogonal matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{\Lambda}$ with

$$\mathbf{Q}^\top\mathbf{A}\mathbf{Q} = \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n). \tag{13.5.1}$$

The diagonal elements λ_i are the eigenvalues of $\mathbf{A}$ and can be any real numbers. Then,

$$Y = \sigma^{-2}\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X} = \sigma^{-2}\boldsymbol{X}^\top\mathbf{Q\Lambda Q}^\top\boldsymbol{X} = \sigma^{-2}\boldsymbol{Z}^\top\mathbf{\Lambda}\boldsymbol{Z} = \sum_{i=1}^{n}\lambda_i(Z_i/\sigma)^2,$$

where $\boldsymbol{Z} = \mathbf{Q}^\top \boldsymbol{X}$. By Theorem 13.3.1, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{b}, \sigma^2 \mathbf{I}_n)$, where $\boldsymbol{b} = \mathbf{Q}^\top \boldsymbol{\mu}$. The MGF of Y is

$$\begin{aligned} M_Y(t) = \mathbb{E}\left[\exp\left(t\sum_{i=1}^n \lambda_i (Z_i/\sigma)^2\right)\right] &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(t\lambda_i (Z_i/\sigma)^2\right)\right] \\ &= \prod_{i=1}^n (1-2t\lambda_i)^{-1/2} \exp\left(\frac{t\lambda_i (b_i/\sigma)^2}{1-2t\lambda_i}\right) \\ &= \left(\prod_{i=1}^n (1-2t\lambda_i)^{-1/2}\right) \cdot \exp\left(\sum_{i=1}^n \frac{t\delta_i}{1-2t\lambda_i}\right), \end{aligned}$$

where $\delta_i = \sigma^{-2}\lambda_i b_i^2 = \sigma^{-2}\boldsymbol{b}^\top \mathbf{\Lambda}_i \boldsymbol{b}$, and $\mathbf{\Lambda}_i = \operatorname{diag}(0, \ldots, 0, \lambda_i, 0, \ldots, 0)$. Thus,

$$\sum_{i=1}^n \delta_i = \sigma^{-2}\boldsymbol{b}^\top \sum_{i=1}^n \mathbf{\Lambda}_i \boldsymbol{b} = \sigma^{-2}\boldsymbol{b}^\top \mathbf{\Lambda} \boldsymbol{b} = \sigma^{-2}\boldsymbol{\mu}^\top \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top \boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\mu}^\top \mathbf{A}\boldsymbol{\mu}.$$

"$\Leftarrow$": Given that $\mathbf{A}$ is idempotent of rank r, we have r of λ_i are 1 and the rest are 0, and

$$M_Y(t) = (1-2t)^{-r/2} \exp\left(\frac{t\delta}{1-2t}\right).$$

Therefore, $Y = \sigma^{-2}\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} \sim \chi^2(r, \delta)$.

"$\Rightarrow$": By the unique factorization of polynomials, r of the λ_i are 1 and the rest are 0. □

Theorem 13.5.3

Let $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and let $\mathbf{A}$ be a symmetric matrix. Then $Y = \boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} \sim \chi^2(r)$ iff $\mathbf{A\Sigma}$ is idempotent of rank r.

Proof Let $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and $\boldsymbol{U} = \mathbf{\Sigma}^{1/2}\boldsymbol{Z}$, then $\boldsymbol{X}$ has the distribution as $\boldsymbol{U}$ from Theorem 13.3.1. Thus the distribution of Y is that of $\boldsymbol{Z}^\top \mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}\boldsymbol{Z}$. From Theorem 13.5.2, $Y \sim \chi^2(r)$ is equivalent to $\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$ is idempotent of rank r. Eventually, the results follows from the fact that the eigenvalues of $\mathbf{\Sigma}^{1/2}\mathbf{A}\mathbf{\Sigma}^{1/2}$ are the same as those of $\mathbf{A\Sigma}$. □

Theorem 13.5.4

Suppose that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is positive-definite. Then

$$Q = (\boldsymbol{X} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2(n). \tag{13.5.2}$$

Proof Making the transformation $\boldsymbol{Y} = \mathbf{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}$, we get

$$Q = \boldsymbol{Z}^\top \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{1/2}\boldsymbol{Z} = \boldsymbol{Z}^\top \boldsymbol{Z} = \sum_{i=1}^n Z_i^2.$$

Since the Z_i^2's are independent $\chi^2(1)$ variables, $Q \sim \chi^2(n)$. □

Theorem 13.5.5

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ and let $\mathbf{A}$, $\mathbf{B}$ be symmetric, idempotent matrices. Then $\boldsymbol{X}^\top \mathbf{A} \boldsymbol{X}$ and $\boldsymbol{X}^\top \mathbf{B} \boldsymbol{X}$ are independent iff $\mathbf{AB} = \mathbf{O}$.

Proof Suppose $\mathbf{AB} = \mathbf{O}$. Since $\mathbf{A}$ and $\mathbf{B}$ are symmetric and idempotent, we can write the quadratic forms as $\boldsymbol{X}^\top \mathbf{A} \boldsymbol{X} = \boldsymbol{X}^\top \mathbf{A}^\top \mathbf{A} \boldsymbol{X} = \|\mathbf{A}\boldsymbol{X}\|^2$ and $\boldsymbol{X}^\top \mathbf{B} \boldsymbol{X} = \|\mathbf{B}\boldsymbol{X}\|^2$. By Theorem 13.4.2, $\mathbf{A}\boldsymbol{X}$ and $\mathbf{B}\boldsymbol{X}$ are independent, which implies that the quadratic forms are independent.

Conversely, let $\boldsymbol{X}^\top \mathbf{A} \boldsymbol{X}$ and $\boldsymbol{X}^\top \mathbf{B} \boldsymbol{X}$ be independent. By Theorem 13.5.2, they are chi-square and therefore $\boldsymbol{X}^\top (\mathbf{A} + \mathbf{B}) \boldsymbol{X}$ must be chi-square. Again, by Theorem 13.5.2, $\mathbf{A} + \mathbf{B}$ must be idempotent. Therefore

$$\mathbf{A} + \mathbf{B} = (\mathbf{A} + \mathbf{B})^2 = \mathbf{A}^2 + \mathbf{B}^2 + \mathbf{AB} + \mathbf{BA} = \mathbf{A} + \mathbf{B} + \mathbf{AB} + \mathbf{BA},$$

so that

$$\mathbf{AB} + \mathbf{BA} = \mathbf{O}.$$

Multiplying on the left by $\mathbf{A}$ gives $\mathbf{AB} + \mathbf{ABA} = \mathbf{O}$, while multiplying on the right by $\mathbf{A}$ gives $\mathbf{ABA} + \mathbf{BA} = \mathbf{O}$; hence $\mathbf{AB} = \mathbf{BA} = \mathbf{O}$. □

Theorem 13.5.6

Suppose that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, $\mathbf{A}$ is an $n \times n$ symmetric and idempotent matrix, and $\mathbf{B}$ is an $n \times p$ matrix. Then $\boldsymbol{X}^\top \mathbf{A} \boldsymbol{X}$ and $\mathbf{B}^\top \boldsymbol{X}$ are independent iff $\mathbf{B}^\top \mathbf{A} = \mathbf{O}$.

Proof Assume, without loss of generality, that the first r eigenvalues of $\mathbf{A}$ are unity. Hence there exists an orthogonal matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^\top \mathbf{A} \mathbf{Q} = \begin{bmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}.$$

Let $\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix}$, where $\mathbf{Q}_1$ is an $n \times r$ matrix, then

$$\begin{aligned} \mathbf{I}_n = \mathbf{Q}^\top \mathbf{Q} &= \begin{bmatrix} \mathbf{Q}_1^\top \\ \mathbf{Q}_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1^\top \mathbf{Q}_1 & \mathbf{Q}_1^\top \mathbf{Q}_2 \\ \mathbf{Q}_2^\top \mathbf{Q}_1 & \mathbf{Q}_2^\top \mathbf{Q}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{n-r} \end{bmatrix} \\ &= \mathbf{Q}\mathbf{Q}^\top = \mathbf{Q}_1 \mathbf{Q}_1^\top + \mathbf{Q}_2 \mathbf{Q}_2^\top. \end{aligned}$$

Let $\boldsymbol{Z} = \mathbf{Q}^\top \mathbf{X}$, then $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{Q}^\top \boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$. Thus,

$$\boldsymbol{X}^\top \mathbf{A} \boldsymbol{X} = \boldsymbol{Z}^\top \mathbf{Q}^\top \mathbf{A} \mathbf{Q} \boldsymbol{Z} = \sum_{i=1}^{r} Z_i^2, \quad \mathbf{B}^\top \boldsymbol{X} = \mathbf{B}^\top \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \boldsymbol{Z}.$$

Moreover,

$$\mathbf{B}^\top \mathbf{A} = \mathbf{B}^\top \mathbf{Q} \begin{bmatrix} \mathbf{I}_r & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \mathbf{Q}^\top = \mathbf{B}^\top \mathbf{Q}_1 \mathbf{Q}_1^\top.$$

"$\Leftarrow$": Suppose that $\mathbf{B}^\top\mathbf{A} = \mathbf{O}$, then

$$\mathbf{B}^\top\mathbf{Q}_1 = \mathbf{B}^\top\mathbf{Q}_1\mathbf{Q}_1^\top\mathbf{Q}_1 = \mathbf{B}^\top\mathbf{A}\mathbf{Q}_1 = \mathbf{O}.$$

Hence, $\mathbf{B}^\top\boldsymbol{X} = \mathbf{B}^\top\mathbf{Q}_2\boldsymbol{Z}_2$, where $\boldsymbol{Z}_2 = (Z_{r+1}, \dots, Z_n)^\top$. Since $Z_1, \dots, Z_n$ are independent, $\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}$ and $\mathbf{B}^\top\boldsymbol{X}$ are independent.

"$\Rightarrow$": Suppose $\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}$ and $\mathbf{B}^\top\boldsymbol{X}$ are independent, then we must have $\mathbf{Q}_1 = \mathbf{O}$. □

Theorem 13.5.7: Cochran's Theorem

Let $\mathbf{A}_1, \dots, \mathbf{A}_k$ be $n \times n$ matrices with $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$. Then the following conditions are equivalent:

(i) $\sum_{i=1}^k \operatorname{rank}(\mathbf{A}_i) = n$;

(ii) $\mathbf{A}_i^2 = \mathbf{A}_i$, for $i = 1, \dots, k$;

(iii) $\mathbf{A}_i\mathbf{A}_j = \mathbf{O}$, for $i \neq j$.

Proof (i) $\Rightarrow$ (iii): Let $\mathbf{A}_i = \mathbf{B}_i\mathbf{C}_i$ be a rank factorization, where $\mathbf{B}_i$ is a $n \times r$ matrix and $\mathbf{C}_i$ is a $r \times n$ matrix, for $i = 1, \dots, k$. Then

$$\mathbf{I}_n = \sum_{i=1}^k \mathbf{A}_i = \sum_{i=1}^k \mathbf{B}_i\mathbf{C}_i = \begin{bmatrix}\mathbf{B}_1 & \cdots & \mathbf{B}_k\end{bmatrix}\begin{bmatrix}\mathbf{C}_1 \\ \vdots \\ \mathbf{C}_k\end{bmatrix}.$$

Since $\sum_{i=1}^k \operatorname{rank}(\mathbf{A}_i) = n$, $\begin{bmatrix}\mathbf{B}_1 & \cdots & \mathbf{B}_k\end{bmatrix}$ is a square matrix and therefore

$$\mathbf{I}_n = \begin{bmatrix}\mathbf{C}_1 \\ \vdots \\ \mathbf{C}_k\end{bmatrix}\begin{bmatrix}\mathbf{B}_1 & \cdots & \mathbf{B}_k\end{bmatrix}.$$

Thus for $i \neq j$, $\mathbf{C}_i\mathbf{B}_j = \mathbf{O}$, hence $\mathbf{A}_i\mathbf{A}_j = \mathbf{B}_i\mathbf{C}_i\mathbf{B}_j\mathbf{C}_j = \mathbf{O}$.

(iii) $\Rightarrow$ (ii): Since $\mathbf{A}_i\mathbf{A}_j = \mathbf{O}$ for $i \neq j$, we have

$$\mathbf{A}_j = \mathbf{A}_j\mathbf{I}_n = \mathbf{A}_j\sum_{i=1}^k \mathbf{A}_i = \mathbf{A}_j^2, \quad j = 1, \dots, k.$$

(ii) $\Rightarrow$ (i): Since $\mathbf{A}_i$ is idempotent, $\operatorname{rank}(\mathbf{A}_i) = \operatorname{tr}(\mathbf{A}_i)$. Now

$$\sum_{i=1}^k \operatorname{rank}(\mathbf{A}_i) = \sum_{i=1}^k \operatorname{tr}(\mathbf{A}_i) = \operatorname{tr}\left(\sum_{i=1}^k \mathbf{A}_i\right) = n.$$

That completes the proof. □

Corollary 13.5.1: Cochran's Theorem

Suppose that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, and $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are symmetric matrices with $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$. Let $r_i = \operatorname{rank}(\mathbf{A}_i)$, and $Q_i = \boldsymbol{X}^\top \mathbf{A}_i \boldsymbol{X}$, $i = 1, \ldots, k$. Then the following conditions are equivalent:

(i) $\sum_{i=1}^k r_i = n$;

(ii) $Q_1, \ldots, Q_k$ are independent;

(iii) $\sigma^{-2} Q_i \sim \chi^2(r_i, \delta_i)$, where $\delta_i = \sigma^{-2} \boldsymbol{\mu}^\top \mathbf{A}_i \boldsymbol{\mu}$, for $i = 1, \ldots, k$.

Example 13.5.1

Suppose that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$. Since

$$\begin{aligned} \sum_{i=1}^n X_i^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2, \\ \Leftrightarrow \boldsymbol{X}^\top \mathbf{I}_n \boldsymbol{X} &= \boldsymbol{X}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J}_n \right) \boldsymbol{X} + \boldsymbol{X}^\top \left(\frac{1}{n} \mathbf{J}_n \right) \boldsymbol{X}, \end{aligned}$$

$\operatorname{rank}(\frac{1}{n}\mathbf{J}_n) = 1$ and $\operatorname{rank}(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n) = n - 1$[a], by Cochran's Theorem (Theorem 13.5.1), we have the followings:

(1) $\sum_{i=1}^n (X_i - \bar{X})^2$ and $n\bar{X}^2$ are independent;

(2) $\sigma^{-2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2 \left(n - 1, \left(\sum_{i=1}^n \mu_i^2 - n\bar{\mu}^2 \right) \sigma^{-2} \right)$;

(3) $\sigma^{-2} n \bar{X}^2 \sim \chi^2(1, n\bar{\mu}^2 \sigma^{-2})$.

[a] First of all, we have $\operatorname{rank}(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n) \geq \operatorname{rank}(\mathbf{I}_n) - \operatorname{rank}(\frac{1}{n}\mathbf{J}_n) = n - 1$. On the other hand, since $(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)\mathbf{1}_n = \mathbf{0}_n$, we have $\operatorname{rank}(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n) \leq n - 1$. Thus, $\operatorname{rank}(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n) = n - 1$.

13.6 Multivariate Non-Central t Distribution

Theorem 13.6.1: Multivariate Non-Central t Distribution

If the $n \times 1$ vector $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite, independently of $Z \sim \chi^2(\nu, \lambda)$, then the probability density function of

$$\boldsymbol{t} = \frac{\boldsymbol{X}}{\sqrt{Z/\nu}} \tag{13.6.1}$$

is given by

$$f(\boldsymbol{t}) = \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + 2\lambda\right)\right]}{(\nu\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}} \sum_{i=0}^{\infty}\sum_{k=0}^{\infty} \frac{\lambda^i}{i!}\left(\frac{2}{\nu}\right)^{k/2} \frac{\Gamma[(\nu+2i+n+k)/2]}{\Gamma[(\nu+2i)/2]} \times \frac{1}{k!}\left(\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^k \left(1+\nu^{-1}\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{t}\right)^{-(\nu+2i+n+k)/2}. \tag{13.6.2}$$

If $\boldsymbol{\mu} = \boldsymbol{0}$, then

$$f(\boldsymbol{t}) = \frac{\exp(-\lambda)}{(\nu\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}\frac{\Gamma[(\nu+2i+n)/2]}{\Gamma[(\nu+2i)/2]} \left(1+\nu^{-1}\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{t}\right)^{-(\nu+2i+n)/2}. \tag{13.6.3}$$

If $\lambda = 0$, then

$$f(\boldsymbol{t}) = \frac{\exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)}{(\nu\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}\Gamma(\nu/2)} \sum_{k=0}^{\infty}\left(\frac{2}{\nu}\right)^{k/2}\Gamma[(\nu+n+k)/2] \times \frac{1}{k!}\left(\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^k \left(1+\nu^{-1}\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{t}\right)^{-(\nu+n+k)/2}. \tag{13.6.4}$$

If $\boldsymbol{\mu} = \boldsymbol{0}$ and $\lambda = 0$, then

$$f(\boldsymbol{t}) = \frac{\Gamma[(\nu+n)/2]}{(\nu\pi)^{n/2}\det(\boldsymbol{\Sigma})^{1/2}\Gamma(\nu/2)} \left(1+\nu^{-1}\boldsymbol{t}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{t}\right)^{-(\nu+n)/2}. \tag{13.6.5}$$

Chapter 14

Analysis of Variance (ANOVA)

14.1 Oneway Analysis of Variance

In the oneway analysis of variance (also known as the oneway classification) we assume that data, Y_{ij}, are observed according to a model

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, n_i, \tag{14.1.1}$$

where the θ_i are unknown parameters and the ε_{ij} are error random variable.

Schematically, the data, Y_{ij}, from a oneway ANOVA will look like this:

Treatments				
1	2	3	$\cdots$	k
y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$\vdots$	$\vdots$	y_{3n_3}		$\vdots$
y_{1n_1}	$\vdots$			$\vdots$
	y_{2n_2}			$\vdots$
				y_{kn_k}

Note that we do not assume that there are equal numbers of observations in each treatment group.

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2}_{\mathsf{SST}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{\bar{y}})^2}_{\mathsf{SSB}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}_{\mathsf{SSW}} \tag{14.1.2}$$

Source of Variance	Degrees of freedom	Sum of squares	Mean square	F statistic
Between treatment groups	$k-1$	$\mathsf{SSB} = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{\bar{y}})^2$	$\mathsf{MSB} = \frac{\mathsf{SSB}}{k-1}$	$F = \frac{MSB}{MSW}$
Within treatment groups	$N-k$	$\mathsf{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	$\mathsf{MSW} = \frac{\mathsf{SSW}}{N-k}$	
Total	$N-1$	$\mathsf{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2$		

Chapter 15

Regression and Linear Models

Regression is a method for studying the relationship between a response variable Y and covariates $\boldsymbol{X} = (X_1, \dots, X_p)^\top$. The covariates are also called predictor variables or features. One way to summarize the relationship between $\boldsymbol{X}$ and Y is through the regression function

$$r(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(Y_{\boldsymbol{x}}) = \int y f(y|\boldsymbol{x}) \, \mathrm{d}y. \tag{15.0.1}$$

Our goal is to estimate the regression function $r(\boldsymbol{x})$ from a random sample of size n:

$$(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n) \sim f_{\boldsymbol{X},Y}(\boldsymbol{x}, y), \tag{15.0.2}$$

where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\top$ is the vector of p covariate values for the i-th observation. Let $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \dots, \hat{Y}_n)^\top$ be predicted (fitted) values from the regression, i.e.,

$$\hat{Y}_i = r(\boldsymbol{X}_i) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}_i), \quad i = 1, \dots, n. \tag{15.0.3}$$

Then, the **residual**, also known as **error**, is the deviation predicted from the actual empirical value of data:

$$e_i = Y_i - \hat{Y}_i. \tag{15.0.4}$$

The **residual sum of squares** (RSS), also known as the **sum of squared errors of prediction** (SSE), is the sum of the squares of residuals:

$$\mathsf{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \left\| \boldsymbol{Y} - \hat{\boldsymbol{Y}} \right\|^2. \tag{15.0.5}$$

The **explained sum of squares** (ESS), alternatively known as the **sum of squares due to regression** (SSR), is the sum of the squares of the deviations of the predicted values from the mean value of a response variable:

$$\mathsf{ESS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \left\| \hat{\boldsymbol{Y}} - \bar{Y} \mathbf{1}_n \right\|^2. \tag{15.0.6}$$

The **total sum of squares** (TSS) is the sum of the squares of the difference of the dependent variable and its mean:

$$\begin{aligned} \mathsf{TSS} &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \\ &= \left\| \boldsymbol{Y} - \bar{Y} \mathbf{1}_n \right\|^2 = \boldsymbol{Y}^\top \left(\mathbf{I}_n - n^{-1} \mathbf{J}_n \right) \boldsymbol{Y}. \end{aligned} \tag{15.0.7}$$

The TSS can be decomposed as follows:

$$\begin{aligned}\mathsf{TSS} &= \left\|(\boldsymbol{Y}-\hat{\boldsymbol{Y}})+(\hat{\boldsymbol{Y}}-\bar{Y}\mathbf{1}_n)\right\|^2 = \left\|\boldsymbol{e}+(\hat{\boldsymbol{Y}}-\bar{Y}\mathbf{1}_n)\right\|^2 \\ &= \mathsf{RSS}+\mathsf{ESS}+2\langle \boldsymbol{e}, \hat{\boldsymbol{Y}}-\bar{Y}\mathbf{1}_n\rangle \\ &= \mathsf{RSS}+\mathsf{ESS}+2\hat{\boldsymbol{Y}}^\top\boldsymbol{e}-2\bar{Y}\mathbf{1}_n^\top\boldsymbol{e}. \end{aligned} \tag{15.0.8}$$

15.1 Linear Regression

The **linear model** is a model of the form

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{15.1.1}$$

In a matrix notation, we denote

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ 1 & X_{2,1} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \tag{15.1.2}$$

then the linear model (15.1.1) can be written as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{15.1.3}$$

Definition 15.1.1: Least Squares

The **least squares estimator** for $\boldsymbol{\beta}$ is the estimator that minimize RSS:

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\, \mathsf{RSS}(\boldsymbol{b}) = \underset{\boldsymbol{b}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\, \|\boldsymbol{Y}-\mathbf{X}\boldsymbol{b}\|^2. \tag{15.1.4}$$

Definition 15.1.2: Generalized Inverse

A generalized inverse of an $m \times n$ matrix $\mathbf{A}$ is defined to be any $n \times m$ matrix $\mathbf{A}^-$ that satisfies the condition

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}. \tag{15.1.5}$$

Taking the transpose of (15.1.5), we have

$$\mathbf{A}^\top(\mathbf{A}^-)^\top\mathbf{A}^\top = \mathbf{A}^\top, \tag{15.1.6}$$

so that $(\mathbf{A}^-)^\top$ is a generalized inverse of $\mathbf{A}^\top$.

Theorem 15.1.1

Let $\mathbf{A}$ be an $n \times m$ matrix, and $\Omega = \mathcal{C}(\mathbf{A}) = \{\boldsymbol{y} : \ \boldsymbol{y} = \mathbf{A}\boldsymbol{x}, \ \forall\ \boldsymbol{x}\}$ is the column space of $\mathbf{A}$. Then $\mathbf{P}_\Omega = \mathbf{A}\left(\mathbf{A}^\top\mathbf{A}\right)^-\mathbf{A}^\top$ is the unique orthogonal projection matrix which project any n-vector $\boldsymbol{y}$ onto Ω, namely,

$$\mathbf{P}_\Omega\boldsymbol{y} \in \Omega, \quad (\mathbf{I}_n - \mathbf{P}_\Omega)\boldsymbol{y} \in \Omega^\perp = \ker\left(\mathbf{A}^\top\right) = \left\{\boldsymbol{x} : \ \mathbf{A}^\top\boldsymbol{x} = \mathbf{0}\right\}. \tag{15.1.7}$$

Here, $\left(\mathbf{A}^\top\mathbf{A}\right)^-$ is any generalized inverse of $\mathbf{A}^\top\mathbf{A}$.

Note $\mathbf{P}_\Omega = \mathbf{P}_\Omega^\top$ and $\mathbf{P}_\Omega^2 = \mathbf{P}_\Omega$. By the Definition 15.1.2,

$$\mathbf{A}^\top\mathbf{P}_\Omega = \mathbf{A}^\top. \tag{15.1.8}$$

Taking the transpose of (15.1.8) gives

$$\mathbf{P}_\Omega\mathbf{A} = \mathbf{A}. \tag{15.1.9}$$

Let $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, then we minimize $\|\boldsymbol{Y} - \boldsymbol{\theta}\|^2$ with respect to $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X}) = \Omega$. If we let $\boldsymbol{\theta}$ vary in Ω, $\|\boldsymbol{Y} - \boldsymbol{\theta}\|^2$ will be minimum for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ when $(\boldsymbol{Y} - \hat{\boldsymbol{\theta}}) \perp \Omega$. This is obvious geometrically, and it is readily proved algebraically as follows.

We first note that $\hat{\boldsymbol{\theta}}$ can be obtained via a symmetric idempotent (projection) matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^-\mathbf{X}^\top$, namely $\hat{\boldsymbol{\theta}} = \mathbf{H}\boldsymbol{Y} \in \Omega$, and thus $\boldsymbol{Y} - \hat{\boldsymbol{\theta}} \perp \Omega$. Then

$$\begin{aligned}
\|\boldsymbol{Y} - \boldsymbol{\theta}\|^2 &= \left\|(\boldsymbol{Y} - \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right\|^2 \\
&= \left\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\right\|^2 + \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 + 2(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})^\top(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&= \left\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\right\|^2 + \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2 \\
&\geq \left\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\right\|^2,
\end{aligned}$$

with equality iff $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Now we solve $\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta}$ for $\boldsymbol{\beta}$. Since $\hat{\boldsymbol{\theta}} = \mathbf{H}\boldsymbol{Y}$,

$$\mathbf{X}\boldsymbol{\beta} = \hat{\boldsymbol{\theta}} = \mathbf{H}\boldsymbol{Y} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^-\mathbf{X}^\top\boldsymbol{Y},$$

which implies that any generalized inverse of $\mathbf{X}^\top\mathbf{X}$ gives a solution

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^-\mathbf{X}^\top\boldsymbol{Y}. \tag{15.1.10}$$

Then, the fitted values are given by

$$\hat{\boldsymbol{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\boldsymbol{Y}. \tag{15.1.11}$$

The residuals are given by

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{Y}, \tag{15.1.12}$$

hence

$$\hat{\boldsymbol{Y}}^\top\boldsymbol{e} = \boldsymbol{Y}^\top\mathbf{H}(\mathbf{I}_n - \mathbf{H})\boldsymbol{Y} = \mathbf{0}. \tag{15.1.13}$$

The minimum RSS is given by

$$\mathsf{RSS}(\hat{\boldsymbol{\beta}}) = \boldsymbol{e}^\top\boldsymbol{e} = \boldsymbol{Y}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{Y} = \boldsymbol{Y}^\top\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^\top\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}. \tag{15.1.14}$$

As $\hat{\boldsymbol{\theta}} = \mathbf{X}\boldsymbol{\beta}$ is unique, we note that $\hat{\boldsymbol{Y}}$, $\boldsymbol{e}$, and RSS are unique, irrespective of the rank of $\mathbf{X}$.

Properties of Least Squares Estimators

Let the columns of $\mathbf{X}$ be linearly independent. As $\mathbf{X}$ has rank $p+1$, $\mathbf{X}^\top\mathbf{X}$ is positive-definite and therefore nonsingular.

Theorem 15.1.2

If $\mathbf{X}$ has full rank, then

$$\mathbf{1}_n^\top \boldsymbol{e} = \sum_{i=1}^n e_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) = 0. \tag{15.1.15}$$

and

$$\mathsf{TSS}(\hat{\boldsymbol{\beta}}) = \mathsf{ESS}(\hat{\boldsymbol{\beta}}) + \mathsf{RSS}(\hat{\boldsymbol{\beta}}). \tag{15.1.16}$$

Therefore,

$$\mathsf{TSS}(\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^\top \left(\mathbf{I}_n - n^{-1}\mathbf{J}_n\right) \boldsymbol{Y}, \tag{15.1.17}$$

$$\mathsf{RSS}(\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^\top \left(\mathbf{I}_n - \mathbf{H}\right) \boldsymbol{Y}, \tag{15.1.18}$$

$$\mathsf{ESS}(\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^\top \left(\mathbf{H} - n^{-1}\mathbf{J}_n\right) \boldsymbol{Y}. \tag{15.1.19}$$

Proof Since the first column of $\mathbf{X}$ are all ones and

$$\mathbf{X}^\top \boldsymbol{e} = \mathbf{X}^\top \left[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right] \boldsymbol{Y} = \left(\mathbf{X}^\top - \mathbf{X}^\top\right) \boldsymbol{Y} = \mathbf{0}, \tag{15.1.20}$$

we have $\mathbf{1}_n^\top \boldsymbol{e} = 0$. Substituting $\mathbf{1}_n^\top \boldsymbol{e} = 0$ and (15.1.13) into (15.0.8) gives $\mathsf{TSS} = \mathsf{ESS} + \mathsf{RSS}$. □

If we assume the errors are unbiased, that is

$$\mathbb{E}(\varepsilon_i|\boldsymbol{X}_i) = 0 \quad \Leftrightarrow \quad \mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \tag{15.1.21}$$

then $\mathbb{E}(\boldsymbol{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\theta}$, and

$$\mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}(\boldsymbol{Y}|\mathbf{X}) = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}. \tag{15.1.22}$$

If we assume further that the ε_i are uncorrelated and have the same variance, that is,

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j|\boldsymbol{X}_i, \boldsymbol{X}_j) = \delta_{ij}\sigma^2 \quad \Leftrightarrow \quad \mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}_n, \tag{15.1.23}$$

then by (13.1.7), we have

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) &= \mathrm{Var}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{Y}\middle|\mathbf{X}\right] \\
&= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\,\mathrm{Var}(\boldsymbol{Y}|\mathbf{X})\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.
\end{aligned} \tag{15.1.24}$$

Theorem 15.1.3: Gauss-Markov

Let $\hat{\boldsymbol{\theta}}$ be the least squares estimator of $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, where $\boldsymbol{\theta} = \mathcal{C}(\mathbf{X})$ and $\mathbf{X}$ may not have full rank. Then among the class of linear unbiased estimates of $\boldsymbol{c}^\top\boldsymbol{\theta}$, $\boldsymbol{c}^\top\hat{\boldsymbol{\theta}}$ is the unique estimate with minimum variance. [We say that $\boldsymbol{c}^\top\hat{\boldsymbol{\theta}}$ is the **best linear unbiased estimate**

(BLUE) of $\boldsymbol{c}^\top\boldsymbol{\theta}$.]

Proof Since $\hat{\boldsymbol{\theta}} = \mathbf{H}\boldsymbol{Y}$ and $\mathbf{H}\boldsymbol{\theta} = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\theta}$, we have, for all $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X})$,

$$\mathbb{E}\left(\boldsymbol{c}^\top\hat{\boldsymbol{\theta}}\middle|\mathbf{X}\right) = \boldsymbol{c}^\top\mathbf{X}\mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{c}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{c}^\top\boldsymbol{\theta}.$$

Hence, $\boldsymbol{c}^\top\hat{\boldsymbol{\theta}}$ is a linear unbiased estimator of $\boldsymbol{c}^\top\boldsymbol{\theta}$.

Let $\boldsymbol{d}^\top\boldsymbol{Y}$ be any other linear unbiased estimator of $\boldsymbol{c}^\top\boldsymbol{\theta}$, then $\boldsymbol{c}^\top\boldsymbol{\theta} = \mathbb{E}\left(\boldsymbol{d}^\top\boldsymbol{Y}\middle|\mathbf{X}\right) = \boldsymbol{d}^\top\boldsymbol{\theta}$, or $(\boldsymbol{c}-\boldsymbol{d})^\top\boldsymbol{\theta} = \mathbf{0}$, so that $\boldsymbol{c}-\boldsymbol{d} \perp \mathcal{C}(\mathbf{X})$. Therefore, $\mathbf{H}(\boldsymbol{c}-\boldsymbol{d}) = \mathbf{0}$. Now, since $\boldsymbol{c}^\top\hat{\boldsymbol{\theta}} = \boldsymbol{c}^\top\mathbf{H}\boldsymbol{Y} = \boldsymbol{c}^\top\mathbf{H}^\top\boldsymbol{Y} = (\mathbf{H}\boldsymbol{c})^\top\boldsymbol{Y} = (\mathbf{H}\boldsymbol{d})^\top\boldsymbol{Y}$,

$$\begin{aligned}
\operatorname{Var}\left(\boldsymbol{d}^\top\boldsymbol{Y}\middle|\mathbf{X}\right) - \operatorname{Var}\left(\boldsymbol{c}^\top\hat{\boldsymbol{\theta}}\middle|\mathbf{X}\right) &= \sigma^2\left(\boldsymbol{d}^\top\boldsymbol{d} - \boldsymbol{d}^\top\mathbf{H}^2\boldsymbol{d}\right)\\
&= \sigma^2\boldsymbol{d}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{d}\\
&= \sigma^2\boldsymbol{d}^\top(\mathbf{I}_n - \mathbf{H})^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{d}\\
&= \sigma^2\left\|(\mathbf{I}_n - \mathbf{H})\boldsymbol{d}\right\|^2\\
&\geq 0,
\end{aligned}$$

with equality iff $(\mathbf{I}_n - \mathbf{H})\boldsymbol{d} = \mathbf{0}$, or $\boldsymbol{d} = \mathbf{H}\boldsymbol{d} = \mathbf{H}\boldsymbol{c}$. Hence, $\boldsymbol{c}^\top\hat{\boldsymbol{\theta}}$ has minimum variance and is unique. □

Corollary 15.1.1

If $\mathbf{X}$ has full rank, then $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{a}^\top\boldsymbol{\beta}$ for every vector $\boldsymbol{a}$.

Proof Now $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ implies that $\boldsymbol{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\hat{\boldsymbol{\theta}}$. Hence setting $\boldsymbol{c}^\top = \boldsymbol{a}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ we have that $\boldsymbol{a}^\top\hat{\boldsymbol{\beta}}$ $(= \boldsymbol{c}^\top\hat{\boldsymbol{\theta}})$ is the BLUE of $\boldsymbol{a}^\top\boldsymbol{\beta}$ $(= \boldsymbol{c}^\top\boldsymbol{\theta})$ for every vector $\boldsymbol{a}$. □

Theorem 15.1.4

If $\mathbf{X}$ is an $n \times (p+1)$ matrix of rank r $(r \leq p+1)$, then

$$\tilde{\sigma}^2 = \frac{\mathsf{RSS}}{n-r} = \frac{\boldsymbol{e}^\top\boldsymbol{e}}{n-r} \tag{15.1.25}$$

is an unbiased estimator of σ^2.

Proof Since

$$\begin{aligned}
\mathbb{E}[(n-r)\tilde{\sigma}^2|\mathbf{X}] &= \mathbb{E}\left[\boldsymbol{Y}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{Y}\middle|\mathbf{X}\right] && \text{(by (15.1.14))}\\
&= \sigma^2\operatorname{tr}(\mathbf{I}_n - \mathbf{H}) + \boldsymbol{\theta}^\top(\mathbf{I}_n - \mathbf{H})\boldsymbol{\theta} && \text{(by Theorem 13.2.1)}\\
&= \sigma^2(n-r),
\end{aligned}$$

we have $\mathbb{E}(\tilde{\sigma}^2|\mathbf{X}) = \sigma^2$. □

Distribution

Until now the only assumptions we have made about the ε are (15.1.21) and (15.1.23). If we assume that the ε are also normally distributed, then

$$\varepsilon_i|\boldsymbol{X}_i \sim \mathcal{N}(0,\sigma^2) \quad \Leftrightarrow \quad \boldsymbol{\varepsilon}|\mathbf{X} \sim \mathcal{N}_n(\mathbf{0},\sigma^2\mathbf{I}_n), \tag{15.1.26}$$

and hence $\boldsymbol{Y}|\mathbf{X} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n)$.

Theorem 15.1.5

If $\boldsymbol{Y}|\mathbf{X} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n)$, where $\mathbf{X}$ is an $n \times (p+1)$ matrix of rank $p+1$, then

(i) $\hat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}_{p+1}\left(\boldsymbol{\beta},\sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$.

(ii) $[\mathsf{RSS}(\boldsymbol{\beta}) - \mathsf{RSS}(\hat{\boldsymbol{\beta}})]/\sigma^2 = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\mathbf{X}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\sigma^2 \sim \chi^2(p+1)$.

(iii) $\hat{\boldsymbol{\beta}}$ is independent of $\tilde{\sigma}^2$.

(iv) $\mathsf{RSS}(\hat{\boldsymbol{\beta}})/\sigma^2 = (n-p-1)\tilde{\sigma}^2/\sigma^2 \sim \chi^2(n-p-1)$.

Proof (i) Since $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^-\mathbf{X}^\top\boldsymbol{Y} = \mathbf{C}\boldsymbol{Y}$, say, where $\mathbf{C}$ is a $(p+1)\times n$ matrix such that $\operatorname{rank}\mathbf{C} = \operatorname{rank}\mathbf{X}^\top = \operatorname{rank}\mathbf{X} = p+1$, $\hat{\boldsymbol{\beta}}$ has a multivariate normal distribution (Theorem 13.3.1). In particular, from equations (15.1.22) and (15.1.24), we have (i).

(ii) Since

$$\begin{aligned}
\mathsf{RSS}(\boldsymbol{\beta}) &= \|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 \\
&= \left\|(\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) + \mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\|^2 \\
&= \left\|\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2 + 2\langle \boldsymbol{e},\hat{\boldsymbol{Y}}-\mathbf{X}\boldsymbol{\beta}\rangle + \left\|\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\|^2 \\
&= \mathsf{RSS}(\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\mathbf{X}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}), \qquad \text{[by (15.1.13) and (15.1.20)]}
\end{aligned}$$

and $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\mathbf{X}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\sigma^2 = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top[\operatorname{Var}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$, the result follows from (i) and Theorem 13.5.4.

(iii) Since $(\mathbf{I}_n-\mathbf{H})\boldsymbol{y} \in \ker\left(\mathbf{X}^\top\right)$ for any n-vector $\boldsymbol{y}$, $\mathbf{X}^\top(\mathbf{I}_n-\mathbf{H}) = \mathbf{O}$, then

$$\begin{aligned}
\operatorname{Cov}(\hat{\boldsymbol{\beta}},\boldsymbol{e}) &= \operatorname{Cov}(\hat{\boldsymbol{\beta}},\boldsymbol{Y}-\hat{\boldsymbol{Y}}) = \operatorname{Cov}(\hat{\boldsymbol{\beta}},\boldsymbol{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\
&= \operatorname{Cov}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{Y},(\mathbf{I}_n-\mathbf{H})\boldsymbol{Y}\right] \\
&= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\operatorname{Var}(\boldsymbol{Y})(\mathbf{I}_n-\mathbf{H})^\top \\
&= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathbf{I}_n-\mathbf{H}) \\
&= \mathbf{O}.
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}$ is independent of $\boldsymbol{e}$, $\mathsf{RSS}(\hat{\boldsymbol{\beta}}) = \|\boldsymbol{e}\|^2$ and $\tilde{\sigma}^2$.

(iv) It follows from (15.1.18), Cochran's Theorem 13.5.1 and $\mathbf{X}^\top(\mathbf{I}_n-\mathbf{H}) = \mathbf{O}$. □

Theorem 15.1.6

The $\gamma \times 100\%$ confidence intervals for the β_j's and σ^2 are give by

1. $\left(\hat{\beta}_j - t_{(1+\gamma)/2}\sqrt{\tilde{\sigma}^2\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]_{jj}}, \ \hat{\beta}_j + t_{(1+\gamma)/2}\sqrt{\tilde{\sigma}^2\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]_{jj}}\right)$;
2. $\left(\dfrac{(n-p-1)\tilde{\sigma}^2}{\chi^2_{(1+\gamma)/2}(n-p-1)}, \dfrac{(n-p-1)\tilde{\sigma}^2}{\chi^2_{(1-\gamma)/2}(n-p-1)}\right)$.

Theorem 15.1.7

1. A $\gamma \times 100\%$ confidence region for $\beta_0, \ldots, \beta_p$ is given by the set of solutions to the inequality

$$\mathsf{RSS}(\boldsymbol{\beta}) \leq \mathsf{RSS}(\hat{\boldsymbol{\beta}})\left[1 + \frac{p+1}{n-p-1} f_\gamma(p+1, n-p-1)\right]. \tag{15.1.27}$$

2. A size α test of $H_0 : \beta_m = \cdots = \beta_p = 0$, where $0 \leq m \leq p$, would reject H_0 if

$$\frac{[\mathsf{RSS}(\hat{\boldsymbol{\beta}}_0) - \mathsf{RSS}(\hat{\boldsymbol{\beta}})]/(p-m+1)}{\mathsf{RSS}(\hat{\boldsymbol{\beta}})/(n-p-1)} > f_{1-\alpha}(p-m+1, n-p-1), \tag{15.1.28}$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{Y}$ and $\hat{\sigma}^2 = \mathsf{RSS}(\hat{\boldsymbol{\beta}})/n$ are the MLEs over the full parameter space $\{(\beta_0, \ldots, \beta_p, \sigma^2) : -\infty < \beta_j < \infty, \sigma^2 > 0\}$. $\hat{\boldsymbol{\beta}}_0 = \left(\mathbf{X}_0^\top\mathbf{X}_0\right)^{-1}\mathbf{X}_0^\top\boldsymbol{Y}$ and $\hat{\sigma}_0^2 = \mathsf{RSS}(\hat{\boldsymbol{\beta}}_0)/n$ are the MLEs over the subset of parameter space such that $\beta_m = \cdots = \beta_p = 0$. Here, $\mathbf{X}_0$ is the $n \times m$ matrix consisting of the first m columns of $\mathbf{X}$.

15.2 Maximum Likelihood Estimation

Assuming normality, the likelihood function for the full-rank regression model is the probability density function of $\boldsymbol{Y}$:

$$L(\boldsymbol{\beta}, \sigma^2) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2\right). \tag{15.2.1}$$

Let $l(\boldsymbol{\beta}, v) = \log L(\boldsymbol{\beta}, \sigma^2)$, where $v = \sigma^2$. Then, ignoring constants, we have

$$l(\boldsymbol{\beta}, v) = -\frac{n}{2}\log v - \frac{1}{2v}\left\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2. \tag{15.2.2}$$

Taking the derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$ and v give

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= -\frac{1}{2v}\left(-2\mathbf{X}^\top\boldsymbol{y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\right), \\ \frac{\partial l}{\partial v} &= -\frac{n}{2v} + \frac{1}{2v^2}\left\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2.\end{aligned}$$

Setting $\partial l/\partial \boldsymbol{\beta} = \mathbf{0}$ and $\partial l/\partial v = v$ gives the MLEs:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{Y}, \quad \hat{\sigma}^2 = \frac{1}{n}\left\|\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2. \tag{15.2.3}$$

Chapter 16

Nonparametric Inference

16.1 The Empirical Distribution Function

Definition 16.1.1

Let $X_1, \ldots, X_n$ be a random sample of size n from $f(x)$. The empirical CDF is

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \begin{cases} 0, & x < X_{1:n}, \\ i/n, & X_{i:n} \le x < X_{i+1:n}, \\ 1, & X_{n:n} \le x. \end{cases} \tag{16.1.1}$$

The empirical PDF is

$$\hat{f}(x) = \begin{cases} 1/n, & \text{if } x \in \{X_1, \ldots, X_n\}, \\ 0, & \text{if } x \notin \{X_1, \ldots, X_n\}. \end{cases} \tag{16.1.2}$$

Theorem 16.1.1

At any fixed value of x,

(i) $n\hat{F}_n(x) \sim \mathsf{BIN}(n, F(x))$;

(ii) $\mathbb{E}[\hat{F}_n(x)] = F(x)$;

(iii) $\mathrm{Var}[\hat{F}_n(x)] = F(x)[1 - F(x)]/n$;

(iv) $\hat{F}_n(x) \xrightarrow{p} F(x)$.

Proof At any fixed value of x, because $X_i \sim f(x)$, $\mathbb{P}(X_i \le x) = F(x)$, then $I(X_i \le x) \sim \mathsf{Bin}(1, F(x))$. Therefore, we have (i), (ii) and (iii). Furthermore, using Chebyshev's inequality (2.4.9) we have

$$\mathbb{P}\left(\left|\hat{F}_n(x) - F(x)\right| \ge \varepsilon\right) \le \frac{1}{n\varepsilon^2} F(x)(1 - F(x)),$$

which implies that $\hat{F}_n(x)$ converges in probability to $F(x)$ as $n \to \infty$, that is (iv). □

Theorem 16.1.2: The Glivenko-Cantelli Theorem

Let $X_1, \dots, X_n \sim f(x)$, then

$$\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| \xrightarrow{a.s} 0. \tag{16.1.3}$$

Theorem 16.1.3: The Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality

Let $X_1, \dots, X_n \sim f(x)$, then

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left[\hat{F}_n(x) - F(x)\right] > \varepsilon\right) \le e^{-2n\varepsilon^2}, \quad \forall\, \varepsilon > \sqrt{\frac{1}{2n}\ln 2}, \tag{16.1.4}$$

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| > \varepsilon\right) \le 2e^{-2n\varepsilon^2}, \quad \forall\, \varepsilon > 0. \tag{16.1.5}$$

A confidence band for $\hat{F}_n$

Using the DKW inequality we can get a confidence band for $\hat{F}_n$. Rewriting DKW we get

$$\mathbb{P}\left(\left|\hat{F}_n(x) - F(x)\right| < \varepsilon \text{ for all } x\right) \ge 1 - 2e^{-2n\varepsilon^2}.$$

Equating $\alpha = 2e^{-2n\varepsilon_n^2}$, which implies $\varepsilon_n = \sqrt{\frac{1}{2n}\ln\frac{2}{\alpha}}$, we get

$$\mathbb{P}\left(\left|\hat{F}_n(x) - F(x)\right| < \varepsilon_n \text{ for all } x\right) \ge 1 - \alpha.$$

Taking into consideration that $F(x) \in [0, 1]$ we can get a slightly more refined result. Define

$$L(x) \triangleq \max\{\hat{F}_n - \varepsilon_n, 0\}, \tag{16.1.6a}$$

$$U(x) \triangleq \min\{\hat{F}_n + \varepsilon_n, 1\}. \tag{16.1.6b}$$

Then, for any CDF F and all n

$$\mathbb{P}\left(L(x) \le F(x) \le U(x) \text{ for all } x\right) \ge 1 - \alpha. \tag{16.1.7}$$

Goodness-of-Fit (GoF) Tests Using the Empirical CDF

Let $X_1, \dots, X_n$ be i.i.d. samples from an unknown distribution F. If we wish to infer whether this sample comes from a certain hypothesized distribution F_0 the problem can be cast as the following hypothesis test:

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \ne F_0.$$

Under H_0, the Glivenko-Cantelli theorem tells us that

$$\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F_0(x)\right| \xrightarrow{a.s} 0. \tag{16.1.8}$$

as $n \to \infty$. Hence, any discrepancy measure between $\hat{F}_n$ and F_0 can be used as a reasonable test statistic.

Theorem 16.1.4: Kolmogorov-Smirnov

Let F_0 be a continuous CDF, and let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables with the CDF F_0. Then the Kolmogorov-Smirnov test statistic

$$D_n \triangleq \sup_x \left|\hat{F}_n(x) - F(x)\right| \tag{16.1.9}$$

has the asymptotic distribution:

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}D_n \le x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}. \tag{16.1.10}$$

Theorem 16.1.5: Cramér-Von Mises

Let F_0 be a continuous CDF, and let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables with the CDF F_0. Then the Cramér-Von Mises statistic

$$C_n \triangleq \int \left[\hat{F}_n(x) - F_0(x)\right]^2 \mathrm{d}F_0(x) \tag{16.1.11}$$

has the asymptotic distribution:

$$\lim_{n\to\infty} \mathbb{P}(nC_n \le x) = 1 - \frac{2}{\pi}\sum_{k=1}^{\infty}(-1)^{k+1}\int_{(2k-1)\pi}^{2k\pi} \frac{\exp(-u^2x/2)}{(-u\sin u)^{1/2}}\,\mathrm{d}u. \tag{16.1.12}$$

Theorem 16.1.6: Anderson-Darling

Let F_0 be a continuous CDF, and let $X_1, \dots, X_n$ be a sequence of i.i.d. random variables with the CDF F_0. Then the Anderson-Darling statistic

$$A_n \triangleq \int_{-\infty}^{\infty} \frac{\left[\hat{F}_n(x) - F_0(x)\right]^2}{F_0(x)[1 - F_0(x)]}\,\mathrm{d}F_0(x) \tag{16.1.13}$$

has the asymptotic distribution:

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(nA_n \le z) = \frac{\sqrt{2\pi}}{x}\sum_{j=0}^{\infty}\binom{-\frac{1}{2}}{j}(4j+1)\exp\left(-\frac{(4j+1)^2\pi^2}{8z}\right)&\\ \times\int_0^{\infty}\exp\left(\frac{z}{8(1+w^2)} - \frac{(4j+1)^2\pi^2w^2}{8z}\right)\mathrm{d}w.& \quad (16.1.14)\end{aligned}$$

where $\binom{-\frac{1}{2}}{j} = (-1)^j\Gamma(j+\frac{1}{2})/[\Gamma(\frac{1}{2})j!]$.

Let F_0 be a continuous function. Since $\hat{F}_n$ is piecewise constant and F_0 is a non-decreasing function. Therefore the maximum deviation between $\hat{F}_n$ and F_0 must occur in a neighborhood of the points Y_i, and so the Kolmogorov-Smirnov statistic can be simplified as

$$\begin{aligned} D_n &= \max_{1\le i\le n} \max\left\{\left|\hat{F}_n(X_{i:n}) - F_0(X_{i:n})\right|, \left|\hat{F}_n(X_{i:n}^-) - F_0(X_{i:n}^-)\right|\right\} \\ &= \max_{1\le i\le n} \max\left\{\left|\frac{i}{n} - U_i\right|, \left|\frac{i-1}{n} - U_i\right|\right\}, \end{aligned}$$

where $U_i = F_0(X_{i:n})$ for $i = 1, \dots, n$. The Cramér-Von Mises statistic can be simplified as

$$\begin{aligned} C_n &= \int_{-\infty}^{X_{1:n}} [F_0(x)]^2 \,\mathrm{d}F_0(x) + \sum_{i=1}^{n-1} \int_{X_{i:n}}^{X_{i+1:n}} \left[\frac{i}{n} - F_0(x)\right]^2 \mathrm{d}F_0(x) + \int_{X_{n:n}}^{\infty} [1 - F_0(x)]^2 \,\mathrm{d}F_0(x) \\ &= \frac{1}{3}U_1^3 + \frac{1}{3}\sum_{i=1}^{n-1}\left[\left(U_{i+1} - \frac{i}{n}\right)^3 - \left(U_i - \frac{i}{n}\right)^3\right] - \frac{1}{3}(U_n - 1)^3 \\ &= \frac{1}{3}U_1^3 + \frac{1}{3}\sum_{i=1}^{n-1}\left[U_{i+1}^3 - U_i^3 + \frac{3i^2}{n^2}(U_{i+1} - U_i) - \frac{3i}{n}(U_{i+1}^2 - U_i^2)\right] - \frac{1}{3}(U_n - 1)^3 \\ &= \frac{1}{3}U_1^3 + \frac{1}{3}U_n^3 - \frac{1}{3}U_1^3 + \left(U_n - \sum_{i=1}^{n}\frac{2i-1}{n^2}U_i\right) - \left(U_n^2 - \sum_{i=1}^{n}\frac{1}{n}U_i^2\right) - \frac{1}{3}(U_n - 1)^3 \\ &= \frac{1}{3} + \frac{1}{n}\sum_{i=1}^{n}\left(U_i^2 - \frac{2i-1}{n}U_i\right) \\ &= \frac{1}{3} + \frac{1}{n}\sum_{i=1}^{n}\left(U_i - \frac{2i-1}{2n}\right)^2 - \frac{1}{4n^3}\sum_{i=1}^{n}(2i-1)^2 \qquad \left[\text{by } \sum_{i=1}^{n}(2i-1)^2 = \frac{1}{3}n(4n^2-1)\right] \\ &= \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left(U_i - \frac{2i-1}{2n}\right)^2. \end{aligned} \tag{16.1.14}$$

Similarly, the Anderson-Darling statistic can be simplified as

$$A_n = -1 - \sum_{i=1}^{n}\frac{2i-1}{n^2}\left[\ln U_i + \ln(1 - U_{n+1-i})\right]. \tag{16.1.15}$$

Appendix A

Polynomials

A.1 Partition

Definition A.1.1: Composition & Integer partition

- A **composition** of a positive integer n is a way of writing n as a sum of a sequence of positive integers.
- A **integer partition** of a positive integer n, is a way of writing n as a sum of a non-increasing sequence of positive integers, denoted by $\lambda \vdash n$.

Theorem A.1.1

There are 2^{n-1} compositions of n.

Example A.1.1

- The integer 4 has $2^{4-1} = 2^3 = 8$ compositions:
  ①: 4; ②: $3+1$; ③: $1+3$; ④: $2+2$; ⑤: $2+1+1$; ⑥: $1+2+1$; ⑦: $1+1+2$; ⑧: $1+1+1+1$.
- The integer 4 has five integer partitions:
  ①: 4; ②: $3+1$; ③: $2+2$; ④: $2+1+1$; ⑤: $1+1+1+1$.

Definition A.1.2: Partition a set

A family of sets P is a partition of a set S iff. all of the following conditions hold:

- The family P doesn't contain the empty set, i.e. $\emptyset \notin P$.
- The union of the sets in P is equal to X, i.e. $\bigcup_{\pi \in P} \pi = S$.
- The intersection of any two distinct sets in P is empty, i.e. $A \bigcap B = \emptyset, \forall A, B \in P,\ A \neq B$.

The sets in P are called the **blocks, parts of cells** of the partition.

Example A.1.2

The set $\{1,2,3\}$ has five partitions: ①: $\{\{1\},\{2\},\{3\}\}$; ②: $\{\{1\},\{2,3\}\}$; ③: $\{\{1,2\},\{3\}\}$; ④: $\{\{1,3\},\{2\}\}$; ⑤: $\{\{1,2,3\}\}$.

Definition A.1.3: Factorial

The rising factorial is defined as the polynomial

$$x^{\overline{n}} = x(x+1)\cdots(x+n-1) = \prod_{k=1}^{n}(x+k-1) = \prod_{k=0}^{n-1}(x+k). \tag{A.1.1}$$

The falling factorial is defined as the polynomial

$$(x)_n = x^{\underline{n}} = x(x-1)\cdots(x-n+1) = \prod_{k=1}^{n}(x-k+1) = \prod_{k=0}^{n-1}(x-k). \tag{A.1.2}$$

Definition A.1.4: Stirling number I

The Stirling numbers of the first kind are defined as the coefficients $s(n,k)$ in the expansion of the falling factorial

$$(x)_n = \sum_{k=0}^{n} s(n,k)x^k. \tag{A.1.3}$$

The unsigned Stirling number of the first kind is the number of permutations of n elements with k disjoint cycles and is denoted by

$$c(n,k) = \begin{bmatrix} n \\ k \end{bmatrix}. \tag{A.1.4}$$

Example A.1.3

Of the $3! = 6$ permutations of three elements, there is one permutation with three cycles (the identity permutation $(1)(2)(3)$), three permutations with two cycles ($(1)(23)$, $(12)(3)$, $(13)(2)$) and two permutation with one cycles ((123), (132)). Thus,

$$\begin{bmatrix} 3 \\ 3 \end{bmatrix} = 1, \qquad \begin{bmatrix} 3 \\ 2 \end{bmatrix} = 3, \qquad \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 2. \tag{A.1.5}$$

Theorem A.1.2

The unsigned Stirling number of the first kind obey the recurrence relation:

$$\begin{bmatrix} n+1 \\ k \end{bmatrix} = n\begin{bmatrix} n \\ k \end{bmatrix} + \begin{bmatrix} n \\ k-1 \end{bmatrix} \tag{A.1.6}$$

for $k > 0$ with initial conditions:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = 1, \qquad \begin{bmatrix} n \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ n \end{bmatrix} = 0 \tag{A.1.7}$$

for $n > 0$.

Proof There are two ways to permutate of n elements with k disjoint cycles:

1. Permutate first n elements with $k-1$ disjoint cycles and let the $(n+1)$-th element as a singleton cycle;
2. Permutate first n elements with k disjoint cycles, then put the $(n+1)$-th element to the place behind one of these n elements to form a bigger cycle. □

Theorem A.1.3

The signs of the (signed) Stirling number of the first kind are predictable and depend on the parity of $n-k$, i.e.

$$s(n,k) = (-1)^{n-k} \begin{bmatrix} n \\ k \end{bmatrix}. \tag{A.1.8}$$

Definition A.1.5: Stirling number II

Stirling number of the second kind is the number of ways to partition a set of n elements into k nonempty subsets and is denoted by

$$S(n,k) = \left\{ \begin{matrix} n \\ k \end{matrix} \right\}. \tag{A.1.9}$$

Theorem A.1.4

Stirling number of the second kind obey the recurrence relation:

$$\left\{ \begin{matrix} n+1 \\ k \end{matrix} \right\} = k \left\{ \begin{matrix} n \\ k \end{matrix} \right\} + \left\{ \begin{matrix} n \\ k-1 \end{matrix} \right\} \tag{A.1.10}$$

for $k > 0$ with initial conditions:

$$\left\{ \begin{matrix} 0 \\ 0 \end{matrix} \right\} = 1, \qquad \left\{ \begin{matrix} n \\ 0 \end{matrix} \right\} = \left\{ \begin{matrix} 0 \\ n \end{matrix} \right\} = 0 \tag{A.1.11}$$

for $n > 0$.

Proof There are two ways to partition $n+1$ elements into k nonempty subsets:

1. Partition first n elements into $k-1$ nonempty subsets and let the $(n+1)$-th element as a singleton subset;
2. Partition first n elements into k nonempty subsets, then put the $(n+1)$-th element into one of these k subsets. □

Theorem A.1.5

Stirling number of the second kind can be calculated as:

$$\left\{ {n \atop k} \right\} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n. \tag{A.1.12}$$

Theorem A.1.6

$$(x)_n = \sum_{k=0}^{n} (-1)^{n-k} \left[{n \atop k} \right] x^k, \tag{A.1.13}$$

$$x^{\overline{n}} = \sum_{k=0}^{n} \left[{n \atop k} \right] x^k, \tag{A.1.14}$$

$$x^n = \sum_{k=0}^{n} \left\{ {n \atop k} \right\} (x)_n, \tag{A.1.15}$$

$$x^n = \sum_{k=0}^{n} (-1)^{n-k} \left\{ {n \atop k} \right\} x^{\overline{n}}. \tag{A.1.16}$$

Definition A.1.6: Bell number

The **Bell numbers** count the possible partitions of a set. The n-th of these number, B_n counts the number of different ways to partition a set that has exactly n elements:

$$B_n = \sum_{k=0}^{n} \left\{ {n \atop k} \right\}. \tag{A.1.17}$$

Theorem A.1.7

The Bell numbers satisfy a recurrence relation:

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k, \qquad B_0 = 1. \tag{A.1.18}$$

Theorem A.1.8

The exponential generating function of the Bell numbers is

$$\sum_{k=0}^{\infty} \frac{B_n}{n!} x^n = e^{e^x - 1}. \tag{A.1.19}$$

A.2 Exponential Bell polynomials

The partial (or incomplete) exponential Bell polynomials[†] are a triangular array of polynomials given by

$$B_{n,k}(x_1,x_2,\dots,x_{n-k+1}) = \sum \frac{n!}{j_1!j_2!\cdots j_{n-k+1}!}\left(\frac{x_1}{1!}\right)^{j_1}\left(\frac{x_2}{2!}\right)^{j_2}\cdots\left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{j_{n-k+1}}, \tag{A.2.1}$$

where the sum is taken over all sequences $j_1, j_2, \dots, j_{n-k+1}$ of non-negative integers such that these two conditions are satisfied:

$$\begin{cases} j_1 + j_2 + \cdots + j_{n-k+1} = k, \\ j_1 + 2j_2 + 3j_3 + \cdots + (n-k+1)j_{n-k+1} = n. \end{cases} \tag{A.2.2}$$

The sum

$$B_n(x_1,\dots,x_n) = \sum_{k=1}^{n} B_{n,k}(x_1,x_2,\dots,x_{n-k+1}) \tag{A.2.3}$$

is called the n-th complete exponential Bell polynomial.

The ordinary Bell polynomials can be expressed in the terms of exponential Bell polynomials:

$$\hat{B}_{n+1}(x_1,\dots,x_{n+1}) = \frac{k!}{n!}\sum_{k=1}^{n} B_{n,k}(1!\cdot x_1, 2!\cdot x_2,\dots,(n-k+1)!\cdot x_{n-k+1}). \tag{A.2.4}$$

Some special values of the Bell polynomial:

$$B_{n,k}(0!,1!,\dots,(n-k)!) = c(n,k) = |s(n,k)| = \left[{n \atop k}\right], \tag{A.2.5}$$

$$B_{n,k}(1,1,\dots,1) = S(n,k) = \left\{{n \atop k}\right\}, \tag{A.2.6}$$

$$B_n(1,1,\dots,1) = \sum_{k=1}^{n} B_{n,k}(1,1,\dots,1) = \sum_{k=1}^{n}\left\{{n \atop k}\right\} = B_n. \tag{A.2.7}$$

Recurrence relations

The complete Bell polynomials can be recurrently defined as

$$B_{n+1}(x_1,\dots,x_{n+1}) = \sum_{i=0}^{n}\binom{n}{i}B_{n-i}(x_1,\dots,x_{n-i})x_{i+1} \tag{A.2.8}$$

with the initial value $B_0 = 1$.

The partial Bell polynomials can also be computed efficiently by a recurrence relation:

$$B_{n,k}(x_1,x_2,\dots,x_{n-k+1}) = \sum_{m=1}^{n-k+1} x_m\binom{n-1}{m-1}B_{n-m,k-1}(x_1,x_2,\dots,x_{n-m-k}), \tag{A.2.9}$$

where

$$\begin{cases} B_{0,0} = 1; \\ B_{n,0} = 0 \quad \text{for } n \geq 1; \\ B_{0,k} = 0 \quad \text{for } k \geq 1. \end{cases} \tag{A.2.10}$$

[†] https://en.wikipedia.org/wiki/Bell_polynomials

Generating functions

The exponential partial Bell polynomials can be defined by the double series expansion of its generating function:

$$\begin{aligned}\Phi(t,u) = \exp\left(u\sum_{j=1}^{\infty} x_j \frac{t^j}{j!}\right) &= \sum_{n,k\geq 0} B_{n,k}(x_1,\ldots,x_{n-k+1})\frac{t^n}{n!}u^k \\ &= 1+\sum_{n=1}^{\infty}\frac{t^n}{n!}\left\{\sum_{k=1}^{n} u^k B_{n,k}(x_1,\ldots,x_{n-k+1})\right\}\end{aligned} \tag{A.2.11}$$

In other words, by what amounts to the same, by the series expansion of the exponential:

$$\frac{1}{k!}\left(\sum_{j=1}^{\infty} x_j \frac{t^j}{j!}\right)^k = \sum_{n=k}^{\infty} B_{n,k}(x_1,\ldots,x_{n-k+1})\frac{t^n}{n!}, \qquad k = 0,1,2,\ldots \tag{A.2.12}$$

The complete exponential Bell polynomial is defined by $\Phi(t,1)$, or in other words:

$$\Phi(t,1) = \exp\left(\sum_{j=1}^{\infty} x_j \frac{t^j}{j!}\right) = \sum_{n=0}^{\infty} B_n(x_1,\ldots,x_n)\frac{t^n}{n!}. \tag{A.2.13}$$

Thus, the n-th complete Bell polynomial is given by

$$B_n(x_1,\ldots,x_n) = \frac{\partial^n}{\partial t^n}\exp\left(\sum_{j=1}^{\infty} x_j \frac{t^j}{j!}\right)\Bigg|_{t=0}. \tag{A.2.14}$$

Inverse relations

If we define

$$y_n = \sum_{k=1}^{n} B_{n,k}(x_1,\ldots,x_{n-k+1}), \tag{A.2.15}$$

then we have the inverse relationship

$$x_n = \sum_{k=1}^{n}(-1)^{k-1}(k-1)!B_{n,k}(y_1,\ldots,y_{n-k+1}). \tag{A.2.16}$$

Faà di Bruno's Formula

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n} f(g(x)) = \sum_{k=1}^{n} f^{(k)}(g(x))B_{n,k}(g'(x),g''(x),\ldots,g^{(n-k+1)}(x)). \tag{A.2.17}$$

A.3 Symmetric polynomials

The power sum symmetric polynomial of degree k in n variables $x_1,\ldots,x_n$ written p_k is the sum of all k-th powers of the variables:

$$p_k(x_1,x_2,\ldots,x_n) = \sum_{i=1}^{n} x_i^k, \qquad k = 0,1,2,\ldots \tag{A.3.1}$$

The elementary symmetric polynomial of degree k in n variables $x_1, \dots, x_n$ written e_k are defined by

$$e_k(x_1, x_2, \dots, x_n) = \begin{cases} 1, & k = 0, \\ \sum_{1 \le j_1 < j_2 < \cdots < j_k \le n} x_{j_1} x_{j_2} \cdots x_{j_k}, & 1 \le k \le n, \\ 0, & k > n. \end{cases} \tag{A.3.2}$$

$$\prod_{i=1}^{n} (\lambda - x_i) = \sum_{k=0}^{n} (-1)^k e_k(x_1, \dots, x_n) \lambda^{n-k}. \tag{A.3.3}$$

$$\prod_{i=1}^{n} \frac{1}{1 - \lambda x_i} = \sum_{k=0}^{\infty} h_k(x_1, \dots, x_n) \lambda^k. \tag{A.3.4}$$

Newton's identities

$$\begin{cases} ke_k = \sum_{i=1}^{k} (-1)^{i-1} e_{k-i} p_i, & n \ge k \ge 1, \\ 0 = \sum_{i=k-n}^{k} (-1)^{i-1} e_{k-i} p_i, & k > n \ge 1. \end{cases} \tag{A.3.5}$$

$$\begin{cases} p_k = (-1)^{k-1} ke_k + \sum_{i=1}^{k-1} (-1)^{k-1+i} e_{k-i} p_i, & n \ge k \ge 1, \\ p_k = \sum_{i=k-n}^{k-1} (-1)^{k-1+i} e_{k-i} p_i, & k > n \ge 1. \end{cases} \tag{A.3.6}$$

Bibliography

BAIN, LEE J & ENGELHARDT, MAX 2000 *Introduction to probability and mathematical statistics*, 2nd edn. Duxbury.

LU, TZON-TZER & SHIOU, SHENG-HUA 2002 Inverses of 2 x 2 block matrices. *Computers & Mathematics with Applications* **43** (1-2), 119–129.

RAHMAN, M. & PEARSON, L. M. 2001 Estimation in two-parameter exponential distributions. *Journal of Statistical Computation and Simulation* **70** (4), 371–386.